An exploratory analysis of combined genome-wide SNP data from several recent studies

Blaise Li

Abstract

The usefulness of a 'total-evidence' approach to human population genetics was assessed through a clustering analysis of combined genome-wide SNP datasets. The combination contained only 3146 SNPs. Detailed examination of the results nonetheless enables the extraction of relevant clues about the history of human populations, some pertaining to events as ancient as the first migration out of Africa. The results are mostly coherent with what is known from history, linguistics, and previous genetic analyses. These promising results suggest that cross-studies data confrontation have the potential to yield interesting new hypotheses about human population history.

Key words: Data combination, Graphical representation, Human populations, Single nucleotide polymorphism

1. Introduction

Let this introduction begin with a disclaimer: I am not a population geneticist, but a phylogeneticist who happens to be interested in human population history. The results presented here should not be considered as scientific claims about human population histories, but only as hypotheses that might deserve further investigation.

In human population genetics, numerous papers have recently been published using genome-wide SNP (Single Nucleotide Polymorphism) data for populations of various places in the world. These papers often represent the data by means of PCA (Principal Component Analysis) plots or clustering bar plots. The details of such graphical representations suggest a variety of interesting hypotheses concerning the relationships between populations. However, it is frustrating to see the data scattered between different studies.

Often, a study would use data from other studies, but typically this would be limited to only a few added populations. Would it not be possible and interesting to go further than just adding the populations necessary to test some specific hypothesis? Do some technical problems prevent the analyses of larger data combinations, involving a wider range of populations. From my experience in phylogeny, I had been made aware of the potential value of so-called 'total-evidence' analyses, where data combination helps extracting relevant information from noisy data. Maybe something interesting could emerge from a total-evidence analysis of these genome-wide SNP datasets.

I quickly noted that gathering the data from the published papers was more difficult than expected. Data from human population genetics studies are not as standardised as those used in phylogenetics. In particular, phylogenetic data is usually stored in a centralised public database (NCBI Genbank) in a standardised format. In human population genetics, it seems that each study has its own policy regarding data availability, and its own way of storing it. In the end, I could obtain the data from the HUGO pan-Asian consortium (2009), Reich et al. (2009) and Bryc et al. (2010), as well as those which are publicly available from the HGDP (Cann et al., 2002; Li et al., 2008) and HapMap (The international HapMap consortium, 2003) projects.

After struggling with the file formats and their different ways of coding the genotypes, I could finally assemble the datasets into a single matrix, free from the infamous A/T and G/C SNPs, and which seemed to produce reasonable results on PCA plots (*i.e.* a consistent placement of similar populations from different datasets).

In the next section, I will describe and comment the results of clustering analyses done with the program `frappe` (Tang et al., 2005), in growing number of clusters (K). For practical reasons, I decided to stop at $K = 16$. The clusters were becoming instable from one value of K to the next. This rendered the detailed examination of the results more difficult, and unreasonably time consuming.

The figures were deposited as a file set on the FigShare repository: `http://dx.doi.org/10.6084/m9.figshare.100442`. The figures will be referenced using their individual `dx.doi.org` URL.

2. Results

2.1. Graphical representation of the results

For each clustering analysis, three kinds of bar plots were generated.
One series represents the profiles (proportions of each cluster) at the indi-

vidual level[1]. The list of clusters are reported below the graph, and for each cluster, the population which has the highest average proportion of this cluster is mentioned. The populations are grouped according to their region, their language family and the alphabetical order of their names.

Another series represents the average profiles of the populations[2]. The populations are grouped according to the geography, the language families, and the profiles similarities.

The last series also represents the average profiles of the populations, but there is one graph for each cluster, and for each graph, the populations are ranked according to their proportion of the corresponding cluster[3].

The colours were chosen based on language families and geography. The language families are the first hierarchical levels of the classification adopted by Lewis (2009)[4].

In the bar plots made at the individual level, an exception to the grouping by geography and language family is made for the populations I labeled 'mixed', which I put in the end. Those populations were sampled in a region not corresponding to their geographical origin or have a well-documented history of admixture. It is of course somewhat arbitrary to decide which populations to put in that separate category, as human population history is made of migration and hybridization. For example, the Hakka and Minnan Chinese from Taiwan are more recent inhabitants of the island than the Ami and Atayal Austronesians. Their migration occurred roughly at the same time as the European and African migrations to America. I could have labelled them as 'mixed', since I have done so with the 'non-native' Americans. There are probably other similar cases; my choices are inevitably biased by my perception of human population history.

Clusters are labelled by numbers. When comparing results obtained with different values of K, to avoid ambiguities, I will often add a subscript to the cluster number indicating the value of K for which it was obtained.

Some clusters are well preserved from one value of K to the next. In the detailed description of the results, when such correspondences are not discussed in the text, they are summarized in a table, using the above-mentioned subscript notation.

The colour attributed to a cluster in the bar plots is determined by the

---

[1] http://dx.doi.org/10.6084/m9.figshare.95764
[2] http://dx.doi.org/10.6084/m9.figshare.95765
[3] http://dx.doi.org/10.6084/m9.figshare.95784
[4] http://www.ethnologue.com/family_index.asp

colour attributed to the population showing the highest proportion of that cluster. This generally helps 'tracking' a cluster across the different values of K, except when populations with similar genetic profiles differ according to their linguistic affiliations. A small differential change in cluster proportions between such populations may then lead to different colours being attributed to 'equivalent' clusters for different values of K. This is the case when the European cluster is either most important in Basque or in Sardinians.

2.2. Detailed results

The detailed review of the results is available in annex (p. 26 and following). It shows how clues about human population history can be extracted through close examination. Readers interested in just having an idea of how this information is extracted are invited to read the comments for the first values of K (up to $K = 5$). More motivated readers may read the rest of the description or even make their own examination of the figures.

2.3. Summary of the results

Average profiles of the populations at $K = 2$: `Frappe_K2_pops.pdf`[5]
At $K = 2$, the separation in 2 clusters differentiates between an 'African' trend (cluster 1) and an 'East Asian' trend (cluster 2).

Average profiles of the populations at $K = 3$: `Frappe_K3_pops.pdf`[6]
At $K = 3$, the 3 trends are 'African' (cluster 1), 'European' (cluster 2) and 'East Asian' (cluster 3).

Average profiles of the populations at $K = 4$: `Frappe_K4_pops.pdf`[7]
At $K = 4$, an 'American' cluster (number 4) is added to the three previous ones: 'African' (number 1), 'European' (number 2) and 'East Asian' (number 3).

Average profiles of the populations at $K = 5$: `Frappe_K5_pops.pdf`[8]
At $K = 5$, there is one cluster for each continent:

- cluster 1, the 'African' cluster (more specifically, 'Sub-Saharan');

- cluster 2, the 'European' cluster;

---

[5] http://dx.doi.org/10.6084/m9.figshare.188
[6] http://dx.doi.org/10.6084/m9.figshare.95713
[7] http://dx.doi.org/10.6084/m9.figshare.189
[8] http://dx.doi.org/10.6084/m9.figshare.190

- cluster 3, the 'Asian' cluster (more specifically, 'East Asian');

- cluster 4, the 'Oceanian' cluster;

- cluster 5, the 'American' cluster.

This result is comparable to what has been already obtained with the HGDP sample (Cann et al., 2002).

Average profiles of the populations at $K = 6$: Frappe_K6_pops.pdf[9]
At $K = 6$, the 'East Asian' cluster 3_5 is split into a 'northern' component (cluster 3_6) and a 'southern' component (cluster 4_6).

Average profiles of the populations at $K = 7$: Frappe_K7_pops.pdf[10]
At $K = 7$, the new cluster that appears, number 2_7, having its highest frequencies in Dravidian populations, and more generally in India and Pakistan, represents a 'South Asian' tendency. This cluster seems to principally replace parts of the 'European' (2_6) and 'Oceanian' (5_6) clusters.

Average profiles of the populations at $K = 8$: Frappe_K8_pops.pdf[11]
At $K = 8$, a 'non-Niger-Congo' cluster (2_8) replaces part of the previous 'African' (1_7) and 'European' (3_7) clusters.

Average profiles of the populations at $K = 9$: Frappe_K9_pops.pdf[12]
At $K = 9$, the 'southern East Asian' cluster which was dominant in Mlabri (6_8) is decomposed in two clusters (6_9 and 7_9). There are now 3 'East Asian' clusters:

- Cluster 4_9 is more present in Altaic, Korean and Japanese populations.

- Cluster 6_9 is more present in Austronesian populations.

- Cluster 7_9 is typical of Malaysian Negritos.

Average profiles of the populations at $K = 10$: Frappe_K10_pops.pdf[13]
At $K = 10$, Mlabri have their profile exclusively composed of cluster 7_{10}, which partly substitutes the 'Austronesian' and 'southern East Asian' clusters 6_9 (then 6_{10}) and 7_9 (then 8_{10}).

---

[9]http://dx.doi.org/10.6084/m9.figshare.191
[10]http://dx.doi.org/10.6084/m9.figshare.192
[11]http://dx.doi.org/10.6084/m9.figshare.193
[12]http://dx.doi.org/10.6084/m9.figshare.194
[13]http://dx.doi.org/10.6084/m9.figshare.195

Average profiles of the populations at $K = 11$: `Frappe_K11_pops.pdf`[14]
At $K = 11$, the 'African' trend is now divided in 3 clusters. A new 'Khoisan-Pygmy' cluster (2_{11}) is added to the previously identified 'general Sub-Saharan' and 'East African-West Asian' cluster.

Average profiles of the populations at $K = 12$: `Frappe_K12_pops.pdf`[15]
At $K = 12$, the 'Khoisan-Pygmy' cluster disappears, and a rearrangement of the 'East Asian' clusters occurs:

- There are 2 'Austronesian' clusters (6_{12} and 7_{12}), one of which (6_{12}) is in fact more specific to the non-Filipino populations of the Philippines. Cluster 7_{12} has a reinforced 'Austronesian' character.

- A 'continental South-East Asian' cluster appears.

- The 'northern East Asian' cluster 4 acquires a more 'maritime' flavour.

- The 'Mlabri-specific' and 'Malaysian Negrito-specific' clusters are maintained.

Average profiles of the populations at $K = 13$: `Frappe_K13_pops.pdf`[16]
At $K = 13$, there are several important changes:

- The 'Khoisan-Pygmy' cluster observed at $K = 11$ reappears.

- A new 'Middle Eastern' cluster (4_{13}) appears.

- The cluster specific to the Negritos from the Philippines (6_{12}) disappears.

Average profiles of the populations at $K = 14$: `Frappe_K14_pops.pdf`[17]
At $K = 14$, the 'Middle Eastern' cluster disappears, but the 'Khoisan-Pygmy' cluster is still there. The Asian clusters are highly reorganized:

- There are two 'Austronesian' clusters. Cluster 7_{14} is dominant in Borneo, Java and the Malaysian peninsula and cluster 8_{14} is dominant in the Philippines.

---

[14]http://dx.doi.org/10.6084/m9.figshare.196
[15]http://dx.doi.org/10.6084/m9.figshare.197
[16]http://dx.doi.org/10.6084/m9.figshare.198
[17]http://dx.doi.org/10.6084/m9.figshare.199

- There is a 'southern East Asian' cluster (11_{14}) predominant in Hmong-Mien and Sino-Tibetan populations.

- There is a cluster specific to the Andamanese and Negritos from the Philippines (12_{14}).

- The 'Indian' (4_{14}), 'northern East Asian' (5_{14}), 'Mlabri-specific' (9_{14}), and 'Malaysian Negrito' (10_{14}) clusters can still be identified.

Average profiles of the populations at $K = 15$: `Frappe_K15_pops.pdf`[18]
At $K = 15$, a 'Middle Eastern' cluster is present, as was the case at $K = 13$. The other clusters correspond to those present at $K = 14$.

Average profiles of the populations at $K = 16$: `Frappe_K16_pops.pdf`[19]
At $K = 16$, the cluster specific to the Andamanese populations again disappears. The 'Austronesian' clusters are reorganized, with the appearance of a cluster specific to the non-Filipino populations of the Philippines (10_{16}), as was the case at $K = 12$. The 'American' cluster is now separated in two:

- Cluster 15_{16} is more present in North America, and is almost absent in the Tupi-speaking populations from the Amazon forest (Surui and Karitiana).

- Cluster 16_{16} is highly dominant in the Tupi, but is also present in the other American populations.

3. Discussion

In this section, I will sometimes use distance trees to compare the profiles of the populations. I will call such trees 'profile trees' (see Materials and Methods, p. 22). It should be noted that these do not aim to represent historical relationships between populations, but only similarities between their clustering profiles[20]. The similarities between clustering profiles are however likely to partially reflect historical relationships, and can therefore be used as an exploratory tool to investigate such relationships.

---

[18]http://dx.doi.org/10.6084/m9.figshare.200
[19]http://dx.doi.org/10.6084/m9.figshare.201
[20]The profile trees will contain clusters of clustering profiles, but it should be clear from the context what type of cluster a sentence is about.

3.1. Correlations with geography

Not surprisingly, like in the original studies of the individual datasets, the compositions of the profiles are mainly correlated with geography. For example, in the profile tree for $K = 16$[21], one can clearly see a cluster containing the populations of Sub-Saharan Africa, one containing the populations of North Africa, Middle East, Europe and Caucasus and one containing almost all populations of Pakistan and India (the exceptions being the Tibeto-Burmese-speaking populations, the Himalayan Pahari and the Hazara, which are closer to the cluster containing the populations of Central, North, and East Asia, the Siddi, which are closer to the Sub-Saharan cluster, and the reciprocal exception are the Indians from Singapore, which cluster with the populations of India).

Within the main clusters, other smaller clusters can be found that reflect geography. For example, the populations of the Lesser Sunda Islands cluster with Papuans and Melanesians.

Geographic structure may also be evidenced within a subset of the populations. For example, in profile trees using populations from west and south Eurasia[22], for most values of K, the populations are disposed along the tree in an order that correlates quite well with a west $\leftrightarrow$ east direction: Europe, Middle East, Caucasus, Pakistan, Kashmir, and the rest of India[23]. The differentiation between Pakistan, Kashmir, and the rest of India parallels the north-Indian / south-Indian opposition evidenced in Reich et al. (2009), but with less details within India. This lack of detail could be due to a much smaller number of SNPs, and also to a less conservative way of selecting populations.

3.2. A note on Negritos and the southern route

As early as $K = 3$, the presence of the 'African' cluster in some populations of South and South-East Asia and Oceania was noticed and interpreted as a possible trace of an old genetic background dating back to early waves of migration out of Africa (see annex, p. 28). Among these populations, Papuans, Melanesians, Andamanese and Negritos from the Philippines and the Malaysian peninsula share the particularity of having a morphology in

---

[21]http://dx.doi.org/10.6084/m9.figshare.216

[22]The trees include the populations of Europe, Caucasus, Middle East, Pakistan (except Hazara), and mainland India (except Pahari and Tibeto-Burmese).

[23]See for example http://dx.doi.org/10.6084/m9.figshare.223.

some points similar to the populations of Africa[24]. This is often interpreted as adaptive convergence, because, from the genetic point of view, these populations have no striking similarities. As we shall see, a closer examination of the genetic data reveals that the overall genetic disparity of these populations hides a few intriguing similarities.

The interpretation of the presence of the 'African' cluster in Oceanian populations and ANLS (Andaman, Negrito, Lesser Sunda) as an 'early wave' signature is reinforced when one considers what happens when the 'Oceanian' cluster appears, at $K = 5$. The 'African' cluster not only decreases in Papuans, Melanesians and in the populations of the geographically close Lesser Sunda Islands, but also in the more remote Andamanese and Negritos from the Malaysian peninsula and from the Philippines, while the decrease is much lower in populations of recent African ancestry (see annex, p. 32). This sharing of profile co-variation by scattered populations is best explained by a shared ancient genetic background, dating to a time when the sea level was lower, than by more recent population migrations. Indeed, contrary to other populations of maritime South-East Asia that are well known for their mastery of navigation, Andamanese and Negritos from the Malaysian peninsula and from the Philippines are land-bound hunter-gatherers. But their lifestyle could of course have changed: The case of Mlabri suggests that a 'reversion' to a hunter-gatherer lifestyle may happen (Oota et al., 2005).

At $K = 11$ another interesting observation arises from the appearance of a cluster dominant in San and Pygmies. First, this shows that Khoisan and Pygmies, all traditionally hunter-gatherers, share not only a mode of subsistence, but also some genetic characteristics. Since they are scattered in various places of Sub-Saharan Africa, this could be interpreted as shared ancestry, dating before the spread of the Bantu populations. A less visible consequence of the appearance of the 'Khoisan-Pygmy' cluster is a differential split of the 'African ancestry' of populations outside Africa into the different 'African' components. The portion of putative African ancestry which is represented by the 'Khoisan-Pygmy' cluster is higher in ANLS than in the populations of recent African ancestry (see annex, p. 44).

It should be also noted that when the 'Khoisan-Pygmy' cluster disappears at $K = 12$, the 'Austronesian' cluster is split in two, one of the resulting clusters (6_{12}) having its highest proportion in the Negritos from the Philip-

---

[24]This morphological particularity led the Spanish to use the term 'Negrito' for some populations of the Philippines. This term is also used for the hunter-gatherer populations of the Malaysian peninsula, and sometimes also for the Andamanese populations.

pines Mamanwa, Ati, Ayta and Agta[25]. This cluster disappears at $K = 13$, while the 'Khoisan-Pygmy' cluster reappears. These switches between the presence of one or the other cluster suggests that some aspect of the genetic composition of the Negritos from the Philippines can be either accounted for by the presence of a 'Khoisan-Pygmy' cluster or by a more specific cluster.

The particularity of the African ancestry of ANLS populations can also be evidenced by PCA (Principal Component Analysis). The `smartpca` program of the EIGENSOFT package (Patterson et al., 2006) allows the determination of the principal components using only a subset of the analyzed populations (option `-w`). I used a selection of Sub-Saharan populations (including Pygmies and San, but excluding the atypical Maasai, Luhya and Fulani) to determine the principal components, and then generated the PCA plot of the populations of interest using the first two principal components. The first component differentiates between a 'Khoisan-Pygmy' side and a 'general Sub-Saharan' side. The second principal component reveals the disparity between San, Biaka and Mbuti. Plotting each individual does not allow to see a clear trend, but representing the populations using the averages of the coordinates of their individuals does[26].

The populations with recent known or possible African ancestry tend to be situated on the 'general Sub-Saharan' side, while ANLS populations and Papuans (who could also bear the genetic traces of the first migrants out of Africa) occupy a more intermediate position, as do the south-eastern Bantu populations (who have received genetic input from Khoisan populations). The principal component that differentiates between Khoisan and Pygmy, on one side, and other Sub-Saharan populations on the other side, also differentiates between ANLS and Papuans on one side, and populations of recent African ancestry on the other side.

These observations suggest that (if the 'early wave' origin of the African component detected in ANLS is accepted) the early out-of-Africa migrants did hold a share of the African genetic diversity more similar to that retained by Khoisan and Pygmies than that retained by other African populations (see annex, p. 44). Another fact that supports this hypothesis is that the morphological characteristics shared by some ANLS populations with Khoisan and Pygmies are not only general features of African populations such as skin colour and hair type, but also more specific characteristics, like short stature. Quite interestingly, Onge and Pygmy women are even subject to steatopygia, an uncommon physical feature for which Khoisan are well known. It would

---

[25]See http://dx.doi.org/10.6084/m9.figshare.301.
[26]See http://dx.doi.org/10.6084/m9.figshare.23398.

be interesting to test whether these shared characteristics could be inherited from a common ancestor, rather than simply be adaptive convergences.

3.3. Austronesian affinities

The PASNP data for Asian populations (HUGO pan-Asian consortium, 2009) used in the present work concern a large number of populations and a relatively smaller number of SNPs than the other datasets. Since the dataset combination consisted in an union of the populations and in an intersection of the SNPs, the assembled dataset probably carries more detailed information for Asian populations than for the other parts of the world. In particular, this permitted marked distinctions between Austronesians. Among these populations, for high values of K, the following groups can be distinguished[27]:

- populations of the Lesser Sunda Islands;

- Iraya and Negritos from the Philippines;

- Mentawai, Toraja, Manobo, Filipinos and Taiwanese (the latter two being more often grouped together);

- populations of the Malaysian peninsula, Sumatra (except Mentawai), Java and Borneo, with the following subgroups:

  - Batak and Malays;

  - Temuans and populations of Java and Borneo.

Below $K = 12$, the cluster containing the populations of the Lesser Sunda Islands is included in the cluster containing the Negritos from the Philippines, and Iraya tend to form a more distant branch[28]. Below $K = 7$, the clusters tend to disaggregate[29].

On profile trees including Tai-Kadai and Austronesian populations, Tai-Kadai tend to cluster with Taiwanese and Filipinos. This is approximately the case from $K = 2$ to $K = 5$[30], and exact for $K = 6$ to $K = 11$ and at $K = 13$[31], but with a growing branch length for the Tai-Kadai sub-group as

---

[27]See for example http://dx.doi.org/10.6084/m9.figshare.244.
[28]See for example http://dx.doi.org/10.6084/m9.figshare.241.
[29]See for example http://dx.doi.org/10.6084/m9.figshare.235.
[30]See for example http://dx.doi.org/10.6084/m9.figshare.248.
[31]See for example http://dx.doi.org/10.6084/m9.figshare.250.

K increases[32]. At $K = 12$, $K = 14$, $K = 15$ and $K = 16$, Tai-Kadai form a separate cluster[33].

If Tai-Kadai have a part of Austronesian ancestry, the profile similarities between Tai-Kadai, Taiwanese and Filipinos suggest that the Austronesian ancestors of Tai-Kadai populations were probably an early offshoot of the Austronesian dispersal (hypothesized to have started from Taiwan). This is compatible with the linguistic evidence detailed in Sagart (2004) (see also annex, p. 46). However, in the profile trees including all populations, this relationship between Tai-Kadai and 'basal' Austronesians is obscured by the fact that, depending on the value of K, Tai-Kadai sometimes cluster with Chinese and Hmong-Mien populations[34]. Moreover, Mon-Khmer and JKL (Jinuo, Karen, Lahu) populations sometimes also cluster with Austronesians[35]. For high values of K the non-Mlabri and non-Negrito Mon-Khmer populations tend to cluster with JKL, Temuans and the populations of Java and Borneo[36].

One may regret the absence of Polynesians (easternmost Austronesians), Malagasy (Austronesians who migrated to the west of the Indian Ocean) and Cham (see the discussion concerning the presence of cluster 7_{12} in Cambodians, p. 46 of the annex) populations in the dataset. This would have offered an even better coverage of the diversity of the Austronesian populations.

3.4. Trans-linguistic affinities

A few trans-linguistic clusters repeatedly appear in the profile trees. Besides the above-mentioned grouping of the populations of the Lesser Sunda Islands with Melanesians and Papuans, one should notice the grouping of the Indo-Iranian Hazara with the Altaic Uyghur. This constitutes a strong evidence for attributing Hazara an origin in Central Asia. Another atypical Indo-Iranian population are the Pahari, which group with Tibeto-Burmese Spiti. Their profile similarities probably reflect genetic exchanges between Tibeto-Burmese and Indo-Iranian populations in the Himalayan region (see also annex, p. 28 and p. 30). A third trans-linguistic grouping involving an Indo-Aryan population is that of Sahariya with Munda. It appears repeatedly, and in some trees, these populations also group with Andamanese. It is difficult to tell whether this might be due to some shared ancestry or if this

---

[32]See for example http://dx.doi.org/10.6084/m9.figshare.255.
[33]See for example http://dx.doi.org/10.6084/m9.figshare.258.
[34]See for example http://dx.doi.org/10.6084/m9.figshare.212.
[35]See for example http://dx.doi.org/10.6084/m9.figshare.209.
[36]See for example http://dx.doi.org/10.6084/m9.figshare.215.

is only an effect of convergent hybridization events between similar Asian genetic stocks. Indeed, the grouping of Fulani with African Americans (and sometimes also with the Maasai) suggests that obviously different histories may produce similarities in the profiles.

3.5. Contrasts within a linguistic family

Differences internal to a linguistic group are also revealed by the comparison of profiles. Different groups of Austronesian populations have been discussed earlier. Other conspicuous cases of 'intra-linguistic' differences can be observed. An interesting example is offered by the Sino-Tibetan family. On profile trees including Sino-Tibetan, Hmong-Mien and Tai-Kadai populations, besides the long branch of the Himalayan Spiti, a striking fact is the particularity of the Tibeto-Burmese populations from the Burmese border (JKL). For most values of K, the profile tree is 'linear', with the populations in the following sequence: Spiti, Tibeto-Burmese of east India (Nysha and Aonaga), Tibeto-Burmese of inner south China (Naxi and Yizu), northern Chinese, Tujia, southern Chinese and She, other Hmong-Mien, eastern Tai-Kadai, western Tai-Kadai, JKL[37]. The JKL have thus profiles quite distinct from those of the other Tibeto-Burmese populations, and in particular distinct from Naxi and Aonaga, which were not sampled very far from the Burmese border, but at more northern locations. Karen, Jinuo and Spiti were listed among the 'linguistic outliers' in the original publication of the data (HUGO pan-Asian consortium, 2009, p. 1543). To be also noted on these profile trees is the difference between the She (which have profiles similar to the neighbouring southern Chinese) and the other Hmong-Mien populations (whose profiles are intermediate between southern Chinese and Tai-Kadai profiles).

Less conspicuous intra-linguistic differences can also be detected on the profile trees. For low values of K, Druze appear to have a profile more similar to European populations than to Palestinians and Bedouins sampled in the same region[38]. The Druze community has its origins at the beginning of the 11th century in the multi-ethnic Fatimid empire. Among its founders are people of Persian and Turk origins, and some famous Druze family names suggest Kurd (Jumblatt) or Turk (Arslan) origins. It may thus be hypothesized that a non-Arab genetic contribution explains the small differences observed between the profiles of Druze and those of the two other populations from Middle East.

---

[37]See for example http://dx.doi.org/10.6084/m9.figshare.273.
[38]See for example http://dx.doi.org/10.6084/m9.figshare.204.

3.6. Profiles co-variation patterns

I will suggest here another manner of using the clustering analyses as an exploratory tool. If the clustering profiles of two population 'react' in the same manner when the clusters are reorganised (that is, when K changes), this may be a sign that these populations share a portion of genetic ancestry inherited from a common population. Therefore, besides considering the direct similarities between profiles, it may be useful to also pay attention to recurrent co-variation patterns[39].

For example, some co-variations are observed between the profiles of the populations of Japan, Taiwan and the Philippines:

- When comparing $K = 12$ with $K = 10$, a rank decrease for the 'Indian' cluster 3 was observed in the Philippines, Taiwan and Japan, and a rank increase occurred for Filipinos and Taiwanese Austronesians for the 'northern East Asian' cluster 4, while the contrast between the populations of Japan and the other populations of northern East Asia was reinforced (see annex, p. 45).

- When comparing the situations at $K = 11$ and $K = 13$ increases in the 'Mlabri-specific' and 'Malaysian Negrito' clusters were observed in the Philippines, Taiwan and Japan (see annex, p. 49).

- When comparing the situations at $K = 12$ and $K = 14$, an increase in the 'southern East Asian' cluster was observed in Taiwan, Japan and the Philippines (see annex, p. 52).

It can be noticed in this respect that the Austronesian populations that have the highest proportion of the northern 'East Asian' cluster (which is dominant in Japan) are Filipinos and Taiwanese Austronesians, for all values of K for which this cluster exists (that is, from $K = 6$ and above).

A possible explanation for these observations could be the maritime activity that occurred in historical times in the region, for instance through Ryukyuan traders. This would have eased the sharing of genetic characteristics between the populations of Taiwan, Japan and the Philippines. More recent events can also be invoked, such as the colonization of Taiwan by the Japanese

---

[39]One could even devise some ways of automatically proposing a correspondence between clusters for different values of K, use this to compute vectors of 'derivatives' of the ancestry profiles for the populations, and build distance trees between these vectors, in order to facilitate the detection of such co-variation patterns.

empire or Japanese migrations to the Philippines during the first half of the 20th century.

Another example is that some co-variations are observed between the profiles of Okinawans and of the populations of the Andaman islands:

- When comparing the situations at $K = 11$ and $K = 13$, a simultaneous decrease was observed in the 'Oceanian' cluster for Okinawans and Onge (see annex, p. 50).

- The 'Oceanian' cluster decreased in Andamanese populations at $K = 14$, when the cluster specific to Andamanese populations appeared (12_{14}), and a strong rank decrease was then observed in that cluster for Okinawans (see annex, p. 53).

These correlations could make sense in the light of the fact that both Andamanese and Okinawans have been reported to have a high proportion of Y chromosome haplogroup D (see Hammer et al., 2006, p. 51 and p. 55). This would reflect an ancient genetic background shared by these two populations. It could be interesting in this respect to add Ainu samples to the dataset, in order to have a better picture of the ancient genetic landscape of Japan.

Yet another example of co-variation pattern is the already mentioned switches between the presence of a 'Khoisan-Pygmy' cluster and one specific to the Negritos from the Philippines (see p. 10). These switches concur in suggesting to investigate the possibility that Negrito populations could share some ancient genetic background with Pygmies and Khoisan populations.

4. Conclusions

When the analyses were performed, the data available from the PASNP consortium did only contain autosomal SNPs. The combined dataset does therefore not contain SNPs located in the Y or mitochondrial chromosomes. The results obtained here are thus complementary to what can be inferred from the studies of Y or mtDNA haplogroups.

If the clusters are to be interpreted as ancestry classes, low values of K might reflect inheritance from older ancestral populations than high values of K. Although more accurate for describing similarities between extant populations, bar plots made with high values of K would then be less likely to reflect ancient historical events. By focusing only on one value of K, or on a narrow range, one might miss some clues about population history. I

would therefore suggest that a wide range of values of K be considered when clustering analyses are used as an exploratory tool.

Despite the small number of SNPs in the combined dataset, the clustering bar plots seem to convey a significant amount of relevant information about human population history[40]. Therefore, the practice consisting in combining data at a large geographical scale seems promising and should be tried with an even more diverse population sampling. This 'taxonomical total-evidence' approach (I borrow here vocabulary from phylogenetics) would be facilitated if the data were stored in a central repository, under a standardised format, and could be more powerful with a better SNP overlap between studies.

Although this work probably does not bring many new results in human population history, I enjoyed the experience and hope that my remarks from outside can be useful to the community of human population genetics.

5. Materials and Methods

5.1. Data preparation

The SNP data were obtained from the following sources:

- 'HGDP' (Cann et al., 2002; Li et al., 2008): the Stanford University HGDP-CEPH SNP genotyping data, supplement 1 (1043 samples);

- 'HapMap' (The international HapMap consortium, 2003): draft release 2 for the genome-wide SNP genotyping of the phase 3 samples (1184 samples);

- 'Asia' (HUGO pan-Asian consortium, 2009): the PASNP consortium genotype data (1928 samples, only the autosomal SNPs were included in the present study);

- 'India' (Reich et al., 2009): SNP data for various populations of India, including populations from the Andaman Islands (132 samples);

- 'Africa' (Bryc et al., 2010): SNP data for various populations of Africa (370 samples).

---

[40]Preliminary analyses using one more source in the combination (the data from Xing et al., 2010) indicate that similar clustering patterns are obtained using only 1656 SNPs. See http://dx.doi.org/10.6084/m9.figshare.89584

According to http://www.cephb.fr/common/RosenbergPreprint.pdf, the HGDP samples include related individuals and 13 duplicates, one of which is labelled both as a Hazara and as a Pathan individual. The duplicates were apparently already suppressed from the downloaded dataset, and the bi-labelled individual completely removed. I had to remove the mis-labelled Biaka Pygmy and Japanese individuals reported in that same document. Some of the HapMap samples are grouped in (mother, father, child) triplets. For such samples, the child was removed.

The data for all remaining samples were combined using `python` (http://www.python.org/) scripts, keeping only the SNPs that were present in the five datasets. The format of the source data differed, and it was not always clear how SNP states between 2 datasets compared. PCA analyses using the `smartpca` program (Patterson et al., 2006) did not show obvious inconsistencies when comparing geographically close populations from different datasets. The resulting combined dataset consists in the genotypes of 4025 individuals at 3146 SNPs. The distribution of the SNPs is summarized in the following table:

chromosome	1	2	3	4	5	6	7	8	9	10	11
# of SNPs	262	264	203	222	241	209	175	166	132	178	166

chromosome	12	13	14	15	16	17	18	19	20	11	22
# of SNPs	160	146	115	99	71	74	100	22	76	45	20

Some populations are sampled in more than one dataset, under different names (for example Uyghur in HUGO pan-Asian consortium (2009) and Uygur in Li et al. (2008)). I kept the original names. The populations are thus distinguished in the admixture graphs, but I used only one spelling in the present text. The two samples did not need to be distinguished in the comments, given the high similarity of their clustering profiles.

The following table gives the list of the sampled populations, with the associated linguistic information:

Population	Language group	Language sub-group
Adygei	North-Caucasian	West-Caucasian
African American	Indo-European	Germanic
Agta	Austronesian	Malayo-Polynesian
Alorese	Austronesian	Malayo-Polynesian
Ami	Austronesian	East-Formosan
Aonaga	Sino-Tibetan	Tibeto-Burman
Atayal	Austronesian	Atayalic
Ati	Austronesian	Malayo-Polynesian
Ayta	Austronesian	Malayo-Polynesian

Population	Language group	Language sub-group
Balochi	Indo-European	Indo-Iranian
Bamoun	Niger-Congo	Atlantic-Congo
Bantu NE	Niger-Congo	Atlantic-Congo
Bantu SE Pedi	Niger-Congo	Atlantic-Congo
Bantu SE Sotho	Niger-Congo	Atlantic-Congo
Bantu SE Tswana	Niger-Congo	Atlantic-Congo
Bantu SE Zulu	Niger-Congo	Atlantic-Congo
Bantu SW Herero	Niger-Congo	Atlantic-Congo
Bantu SW Ovambo	Niger-Congo	Atlantic-Congo
Batak Karo	Austronesian	Malayo-Polynesian
Batak Toba	Austronesian	Malayo-Polynesian
Bedouin	Afro-Asiatic	Semitic
Bengali	Indo-European	Indo-Iranian
Bhil	Indo-European	Indo-Iranian
Bhili	Indo-European	Indo-Iranian
Biaka Pygmies	Niger-Congo	Atlantic-Congo
Bidayuh Jagoi	Austronesian	Malayo-Polynesian
Brahui	Dravidian	Northern-Dravidian
Brong	Niger-Congo	Atlantic-Congo
Bulala	Nilo-Saharan	Central-Sudanic
Burusho	Burushaski	Burushaski
Cambodians	Austro-Asiatic	Mon-Khmer
Chenchu	Dravidian	South-Central-Dravidian
Chinese Denver	Sino-Tibetan	Chinese
Chinese Hakka	Sino-Tibetan	Chinese
Chinese Minnan	Sino-Tibetan	Chinese
Colombians	Arawakan	Maipuran
Dai	Tai-Kadai	Kam-Tai
Daur	Altaic	Mongolic
Dayak	Austronesian	Malayo-Polynesian
Druze	Afro-Asiatic	Semitic
European Utah	Indo-European	Germanic
Fang	Niger-Congo	Atlantic-Congo
Filipino Ilocano	Austronesian	Malayo-Polynesian
Filipino Tagalog	Austronesian	Malayo-Polynesian
Filipino Visaya Chabakano	Creole	Spanish-based
French	Indo-European	Italic
French Basque	Basque	Basque
Great Andamanese	Andamanese	Great-Andamanese
Gujarati Houston	Indo-European	Indo-Iranian
Hallaki	Dravidian	Southern-Dravidian
Han	Sino-Tibetan	Chinese

Population	Language group	Language sub-group
Han BJ	Sino-Tibetan	Chinese
Han Cantonese	Sino-Tibetan	Chinese
Han Mandarin	Sino-Tibetan	Chinese
Han Singapore	Sino-Tibetan	Chinese
Hausa	Afro-Asiatic	Chadic
Hazara	Indo-European	Indo-Iranian
Hezhen	Altaic	Tungusic
Hindi	Indo-European	Indo-Iranian
Hmong	Hmong-Mien	Hmongic
Hmong Miao	Hmong-Mien	Hmongic
Htin Mal	Austro-Asiatic	Mon-Khmer
Igbo	Niger-Congo	Atlantic-Congo
Indian Singapore	Dravidian	Southern-Dravidian
Iraya	Austronesian	Malayo-Polynesian
Japanese	Japonic	Japanese
Japanese Tokyo	Japonic	Japanese
Javanese	Austronesian	Malayo-Polynesian
Jiamao	Tai-Kadai	Hlai
Jinuo	Sino-Tibetan	Tibeto-Burman
Kaba	Nilo-Saharan	Central-Sudanic
Kalash	Indo-European	Indo-Iranian
Kambera	Austronesian	Malayo-Polynesian
Kamsali	Dravidian	South-Central-Dravidian
Karen	Sino-Tibetan	Tibeto-Burman
Karitiana	Tupi	Arikem
Kashmiri Pandit	Indo-European	Indo-Iranian
Kharia	Austro-Asiatic	Munda
Kongo	Niger-Congo	Atlantic-Congo
Koreans	Korean	Korean
Kurumba	Dravidian	Southern-Dravidian
Lahu	Sino-Tibetan	Tibeto-Burman
Lamaholot	Austronesian	Malayo-Polynesian
Lawa	Austro-Asiatic	Mon-Khmer
Lembata	Austronesian	Malayo-Polynesian
Lodi	Indo-European	Indo-Iranian
Luhya Kenya	Niger-Congo	Atlantic-Congo
Maasai Kenya	Nilo-Saharan	Eastern-Sudanic
Mada	Afro-Asiatic	Chadic
Madiga	Dravidian	South-Central-Dravidian
Makrani	Indo-European	Indo-Iranian
Mala	Dravidian	South-Central-Dravidian
Malay	Austronesian	Malayo-Polynesian

Population	Language group	Language sub-group
Malay Singapore	Austronesian	Malayo-Polynesian
Mamanwa	Austronesian	Malayo-Polynesian
Mandenka	Niger-Congo	Mande
Manggarai	Austronesian	Malayo-Polynesian
Marathi	Indo-European	Indo-Iranian
Maya	Mayan	Yuacatecan
Mbororo Fulani	Niger-Congo	Atlantic-Congo
Mbuti Pygmies	Nilo-Saharan	Central-Sudanic
Meghawal	Indo-European	Indo-Iranian
Melanesians Naasioi	South-Bougainville	Nasioi
Mentawai	Austronesian	Malayo-Polynesian
Mexican LA	Indo-European	Italic
Miaozu	Hmong-Mien	Hmongic
Minanubu Manobo	Austronesian	Malayo-Polynesian
Mlabri	Austro-Asiatic	Mon-Khmer
Mon	Austro-Asiatic	Mon-Khmer
Mongola	Altaic	Mongolic
Mozabite	Afro-Asiatic	Berber
NAN Melanesian	South-Bougainville	Nasioi
Naidu	Dravidian	South-Central-Dravidian
Naxi	Sino-Tibetan	Tibeto-Burman
Negrito Jehai	Austro-Asiatic	Mon-Khmer
Negrito Kensiu	Austro-Asiatic	Mon-Khmer
North Italian	Indo-European	Italic
Nysha	Sino-Tibetan	Tibeto-Burman
Okinawan	Japonic	Ryukyuan
Onge	Andamanese	South-Andamanese
Orcadian	Indo-European	Germanic
Oroqen	Altaic	Tungusic
Pahari	Indo-European	Indo-Iranian
Palestinian	Afro-Asiatic	Semitic
Palaung	Austro-Asiatic	Mon-Khmer
Papuan	Sepik	Ndu
Pathan	Indo-European	Indo-Iranian
Pima	Uto-Aztecan	Southern-Uto-Aztecan
Plang Blang	Austro-Asiatic	Mon-Khmer
Russian	Indo-European	Slavic
Sahariya	Indo-European	Indo-Iranian
San	Khoisan	Southern-africa
Santhal	Austro-Asiatic	Munda
Sardinian	Indo-European	Italic
Satnami	Indo-European	Indo-Iranian

Population	Language group	Language sub-group
She	Hmong-Mien	Ho-Nte
Siddi	Dravidian	Southern-Dravidian
Sindhi	Indo-European	Indo-Iranian
Spiti	Sino-Tibetan	Tibeto-Burman
Srivastava	Indo-European	Indo-Iranian
Sunda	Austronesian	Malayo-Polynesian
Surui	Tupi	Monde
Tai Khuen	Tai-Kadai	Kam-Tai
Tai Lue	Tai-Kadai	Kam-Tai
Tai Yong	Tai-Kadai	Kam-Tai
Tai Yuan	Tai-Kadai	Kam-Tai
Telugu Kannada	Dravidian	Southern-Dravidian
Temuan	Austronesian	Malayo-Polynesian
Tharu	Indo-European	Indo-Iranian
Toraja	Austronesian	Malayo-Polynesian
Toscani Italia	Indo-European	Italic
Tu	Altaic	Mongolic
Tujia	Sino-Tibetan	Tibeto-Burman
Tuscan	Indo-European	Italic
Uyghur	Altaic	Turkic
Uygur	Altaic	Turkic
Vaish	Indo-European	Indo-Iranian
Velama	Dravidian	South-Central-Dravidian
Vysya	Dravidian	South-Central-Dravidian
Wa	Austro-Asiatic	Mon-Khmer
Xhosa	Niger-Congo	Atlantic-Congo
Xibo	Altaic	Tungusic
Yakut	Altaic	Turkic
Yao Iu Mien	Hmong-Mien	Mienic
Yizu	Sino-Tibetan	Tibeto-Burman
Yoruba	Niger-Congo	Atlantic-Congo
Yoruba Nigeria	Niger-Congo	Atlantic-Congo
Zhuang N	Tai-Kadai	Kam-Tai

The colours of the population names in the above table are those that where used in the graphics. These colours where chosen according to linguistic affiliations and geography. They were used to distinguish the clusters in the bar plots (see below).

5.2. Data analysis and visualization

The combined dataset was analysed using the program `frappe` (Tang et al., 2005, `http://med.stanford.edu/tanglab/software/frappe.html`), with K (number of clusters to use) ranging from 2 to 16. The graphics were produced using a combination of `python` scripts and the TikZ/PGF graphic system (`http://sourceforge.net/projects/pgf/`).

In the bar plots, each cluster was given the colour of the population which had the highest proportion of this cluster, except when this rule would have given the same colour to several clusters. In such cases, the clusters where differentiated by darker or lighter shades of the common colour. The goal of these rules was to enable an automatic colour attribution to the clusters. This was necessary given the large amount of graphics produced. Often (but not always: see p. 4), the resulting colour attribution allows the visual recognition of a cluster across the different values of K.

Profile trees used for the discussion were built, for a given value of K and a given selection of populations, by computing the pairwise χ^2 distances between the vectors representing the average profiles of the populations. The distance matrix was then used to build a tree with `fastme` (Desper and Gascuel, 2002). The trees were plotted using a combination of `python` scripts and the TikZ/PGF graphic system.

Conflicts of Interest

The author declares no conflict of interest.

References

Bryc, K., Auton, A., Nelson, M. R., Oksenberg, J. R., Hauser, S. L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C., Tishkoff, S. A., Bustamante, C. D., 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. P. Natl. Acad. Sci. USA 107 (2), 786–791.
URL http://dx.doi.org/10.1073/pnas.0909559107

Cann, H. M., de Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W. F., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G. B., Friedlaender, J. S., Groot, H., Gurwitz, D., Jekins, T., Herrera, R. J., Huang, X., Kidd, J., Kidd, K. K., Langaney, A., Lin, A. A., Mehdi, S. Q., Parham, P., Piazza, A., Pistillo, M. P., Qian, Y., Shu, Q., Xu, J., Zhu, S., Weber, J. L., Greely, H. T., Feldman, M. W., Thomas, G., Dausset, J., Cavalli-Sforza, L. L., 2002. A human genome diversity cell line panel. Science 296 (5566), 261–262.
URL http://dx.doi.org/10.1126/science.1177074

Desper, R., Gascuel, O., 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9 (5), 687–705.
URL http://dx.doi.org/10.1089/106652702761034136

Hammer, M. F., Karafet, T. M., Park, H., Omoto, K., Harihara, S., Stoneking, M., Horai, S., 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. J. Hum. Genet. 51, 47–58.
URL http://dx.doi.org/10.1007/s10038-005-0322-0

HUGO pan-Asian consortium, 2009. Mapping human genetic diversity in Asia. Science 326 (5959), 1541–1545.
URL http://dx.doi.org/10.1126/science.1177074

Lewis, M. P. (Ed.), 2009. Ethnologue: Languages of the World, sixteenth Edition. SIL International, Dallas, Tex.
URL http://www.ethnologue.com/

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., Cavalli-Sforza, L. L., Myers, R. M., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319 (5866), 1100–1104.
URL http://dx.doi.org/10.1126/science.1153717

Oota, H., Pakendorf, B., Weiss, G., von Haeseler, A., Pookajorn, S., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., Stoneking, M., 2005. Recent origin and cultural reversion of a hunter-gatherer group. PLOS Biol. 3 (3), 536–542.
URL http://dx.doi.org/10.1371/journal.pbio.0030071

Patterson, N., Price, A. L., Reich, D., 2006. Population structure and eigenanalysis. PLOS Genet. 2 (12), 2074–2093.
URL http://dx.doi.org/10.1371/journal.pgen.0020190

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E. D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T. L., Grønnow, B., Meldgaard, M., Andreasen, C., Fedorova, S. A., Osipova, L. P., Higham, T. F. G., Ramsey, C. B., Hansen, T. V. O., Nielsen, F. C., Crawford, M. H., Brunak, S., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., Willerslev, E., 2010. Ancient human genome sequence of an extinct palaeo-Eskimo. Nature 463, 757–762.
URL http://dx.doi.org/10.1038/nature08835

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., Singh, L., 2009. Reconstructing Indian population history. Nature 461, 489–494.
URL http://dx.doi.org/10.1038/nature08365

Sagart, L., 2004. The higher phylogeny of Austronesian and the position of Tai-Kadai. Ocean. Linguist. 43 (2), 411–444.
URL http://dx.doi.org/10.1353/ol.2005.0012

Schuster, S. C., Miller, W., Ratan, A., Tomsho, L. P., Giardine, B., Kasson, L. R., Harris, R. S., Petersen, D. C., Zhao, F., Qi, J., Alkan, C., Kidd, J. M., Sun, Y., Drautz, D. I., Bouffard, P., Muzny, D. M., Reid, J. G., Nazareth, L. V., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N. E., Moorjani, P., Tindall, E. A., Danko, C. G., Teo, W. S., Buboltz, A. M., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A. W., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B. F., Makova, K. D., Nekrutenko, A., Mardis, E. R., Patterson, N., Pringle, T. H., Chiaromonte, F., Mullikin, J. C., Eichler, E. E., Hardison, R. C., Gibbs, R. A., Harkins, T. T., Hayes, V. M., 2010. Complete Khoisan and Banu genomes from southern Africa. Nature 463, 943–947.
URL http://dx.doi.org/10.1038/nature08795

Su, B., Xiao, C., Deka, R., Seielstad, M. T., Kangwanpong, D., Xiao, J., Lu, D., Underhill, P., Cavalli-Sforza, L., Chakraborty, R., Jin, L., 2000. Y chromosome haplotypes reveal prehistorical migrations to the Himalayas. Hum. Genet. 107, 582–590.
URL http://dx.doi.org/10.1007/s004390000406

Tang, H., Peng, J., Wang, P., Risch, N. J., 2005. Estimation of individual admixture: analytical and study design considerations. Genet. Epidemiol. 28 (4), 289–301.
URL http://dx.doi.org/10.1002/gepi.20064

The international HapMap consortium, 2003. The international HapMap project. Nature 426, 789–796.
URL http://dx.doi.org/10.1038/02168

Xing, J., Watkins, W. S., Schlien, A., Walker, E., Huff, C. D., Witherspoon, D. J., Zhang, Y., Simonson, T. S., Weiss, R. B., Schiffman, J. D., Malkin, D., Woodward, S. R., Jorde, L. B., 2010. Toward a more uniform sampling of human genetic diversity: A survey of worldwide populations by high-density genotyping. Genomics 96 (4), 199–210.
URL http://dx.doi.org/10.1016/j.ygeno.2010.07.004

Annex: detailed description of the results

K = 2

Raw results: `Frappe_K2.txt`[41]

Profiles of the individuals: `Frappe_K2.pdf`[42]

Average profiles of the populations: `Frappe_K2_pops.pdf`[43]

Ranked average profiles of the populations: `Frappe_K2_rankings.pdf`[44]

The separation in 2 clusters differentiates between a 'Sub-Saharan' trend (cluster 1) and an 'East Asian' trend (cluster 2).

The most typically 'Sub-Saharan' population is a Bantu population, and the most typically 'East Asian' is an Austronesian population from Taiwan. The Bantu populations are known for having spread over a large part of Sub-Saharan Africa during the last millenia and the Austronesians have done the same in the Pacific and Indian oceans, with a probable origin in Taiwan.

African populations have a large predominance of cluster 1. The Sub-Saharan populations with a noticeable component 2 are the Fulani and the Maasai. The Fulani are West-African nomads whose origins are controversial. It is sometimes proposed that they have migrated from more eastern regions of Africa. The Maasai are an East African population which probably originates from North-East Africa. Unfortunately, the dataset lacks some populations from Sudan or from the Horn of Africa.

The proportion of cluster 1 is partly correlated to distance from Sub-Saharan Africa, with the following gradient:
Sub-Saharan Africa > North Africa > Middle East > Europe > Pakistan > India.

As expected from their African ancestry, Siddi ('African Indians') and African Americans have high cluster 1 proportions.

Cluster 1 is noticeable in populations from America and Oceania. It should be noted that the Oceanians in the dataset are not Austronesians. It could be interesting to add some Polynesian populations to the dataset.
Non-Taiwanese Austronesians in the dataset are not among those presenting the highest proportions of cluster 2. This difference with Taiwanese could be

---

[41]http://dx.doi.org/10.6084/m9.figshare.104
[42]http://dx.doi.org/10.6084/m9.figshare.276
[43]http://dx.doi.org/10.6084/m9.figshare.188
[44]http://dx.doi.org/10.6084/m9.figshare.291

explained by some admixture between Malayo-Polynesians and other populations such as Indians in the maritime territories of South-East Asia. In coherence with this hypothesis is the fact that most continental East and South-East Asian populations (Sino-Tibetans, Tai-Kadai, Hmong-Mien and some Austro-Asiatic) show a very high cluster 2 proportion, like the Taiwanese Austronesians. The exceptions are Mon and Cambodians, two Austro-Asiatic populations of Indochina that have a little more cluster 1 proportion than the others (but their profile is still predominantly composed by cluster 2, and the influence of India has been strong on Indochina too).

Altaic populations show various proportions of cluster 1. In this regard, they differ from Koreans and Japanese, to whom they are sometimes related by linguists. Koreans and Japanese have profiles more similar to Sino-Tibetan populations, i.e. a very low cluster 1 proportion. This low proportion in East Asian populations contrasts with what is observed in American populations. If the ancestry of the latter is to be found somewhere in Asia, it would probably not be from a stem with a profile similar to that of extant East Asians. It should be noted that the sample of American populations does not contain Na-Dene or Eskimo-Aleut speakers. Including the data from Rasmussen et al. (2010) could yield interesting results.

K = 3

Raw results: `Frappe_K3.txt`[45]

Profiles of the individuals: `Frappe_K3.pdf`[46]

Average profiles of the populations: `Frappe_K3_pops.pdf`[47]

Ranked average profiles of the populations: `Frappe_K3_rankings.pdf`[48]

The 3 trends are 'African' (cluster 1), 'European' (cluster 2) and 'East Asian' (cluster 3).

Cluster 1 is overwhelming in Sub-Saharan African populations, except for the two previously noted Fulani and Maasai, which show a significant proportion of cluster 2. Among Bantu-speaking population, north-eastern Bantu and Luhya from Kenya show a little more of cluster 2 than the others (which is not surprising, considering the geographic proximity of these populations

---

[45] http://dx.doi.org/10.6084/m9.figshare.105
[46] http://dx.doi.org/10.6084/m9.figshare.277
[47] http://dx.doi.org/10.6084/m9.figshare.95713
[48] http://dx.doi.org/10.6084/m9.figshare.292

with the Maasai). The same holds for the Nilo-Saharan-speaking Bulala. Cluster 1 is dominant in 'African Indians' (Siddi) and African Americans.

Cluster 1 is important in Mozabites from North Africa, Bedouins and Palestinians from Middle East. Some Mozabite and Bedouin individuals have more than 50% cluster 1.

In places geographically more distant to Africa, cluster 1 is found with an important proportion in some individuals in Makrani and Sindhi, populations from southern Pakistan. This could be explained by admixture with descendants from African slaves or soldiers (Sheedis) that are established in these regions.

Cluster 1 is also noticeable in Oceanian populations, and to various degrees in some populations of maritime South-East Asia:

- Onge and Great Andamanese (from the Andaman islands);

- Jehai and Kensiu (Negritos from Malaysia);

- Kambera, Manggarai, Lamaholot, Lembata and Alorese (from the Lesser Sunda Islands);

- Mamanwa, Agta, Ati and Ayta (Negritos from the Philippines).

I will use the abbreviation ANLS to designate this group of populations: Andaman, Negrito, Lesser Sunda. The presence of cluster 1 in these populations could be a genetic trace of the ancient colonization of these regions by an early wave of migration out of Africa. It would be interesting in this regard to add Australian populations to the data, as Australia is thought to have been reached early in the history of world colonization by modern humans.

Cluster 2 is predominant in populations from North Africa, Middle East, Europe, Pakistan and the Dravidian and Indo-European populations of India. There are however some Indo-European-speaking populations with a somewhat lower cluster 2 proportion. For example, Hazara from northern Pakistan, who have some Altaic origins, and Himalayan populations (Pahari), who live in close contact with Sino-Tibetan populations.

Among populations with a high cluster 2 proportion, those from West and South Europe have the highest proportion. The cluster 2 proportion is slightly lower for populations of the Middle East (who have instead a higher cluster 1 proportion) and for populations in East-Europe and Pakistan (who have a higher cluster 3 proportion). For the populations of India the decrease

in cluster 2 ('compensated' by an increase in cluster 3) continues, with a tendency for Dravidian populations to have a lower cluster 2 proportion than Indo-European populations.

Cluster 2 is important in American populations and in some Altaic populations such as Uyghur and Yakut. As for $K = 2$, American populations are more similar in clustering profile to Altaic populations than to other Asian populations. As noted previously (p. 27), the inclusion of the data from Rasmussen et al. (2010) could be highly interesting, because this study not only had Na-Dene and Eskimo-Aleut samples, but also a fair variety of Siberian populations.

Cluster 2 is also important in the Himalayan Sino-Tibetan populations (Spiti). This observation is coherent with the results from the study of Y chromosomes: Himalayan Sino-Tibetan populations have a high diversity of Y haplotypes, indicating complex ancestry (Su et al., 2000). The high proportion of cluster 2 could for example be explained by an Altaic contribution in Spiti's ancestry. Some admixture with Indo-Europeans is also probable, given the localisation of the sampled population (Jammu and Kashmir).

Similarly to cluster 1, cluster 2 is noticeable in various populations of maritime South-East Asia. It is also noticeable in some populations speaking Austro-Asiatic languages: Kharia and Santhal from India, Cambodians, Mon from Thailand, Kensiu and Jehai from peninsular Malaysia. Admixture with neighbouring Indian populations is highly probable in the case of Kharia and Santal, and the hypothesis of an Indian influence in maritime South-East Asia proposed for $K = 2$ (p. 27) can be invoked again to explain the presence of cluster 2 in the populations of South-East Asia.

Cluster 3 is highly predominant in Hmong-Mien and Tai-Kadai populations, most Sino-Tibetan populations, Koreans, Japanese, and some Austronesian populations: Atayal and Ami (from Taiwan), Bidayuh and Dayak from Borneo, Mentawai (west of Sumatra), Toraja (from Sulawesi), Manobo and Filipinos (from the Philippines). More generally, it is by far the main component in all populations from East and South-East Asia, and constitutes an important part of the clustering profiles of populations from Oceania, America and Central and North Asia. It decreases in favor of cluster 2 following an east > west gradient in populations of India, Pakistan and East Europe.

K = 4

Raw results: `Frappe_K4.txt`[49]

Profiles of the individuals: `Frappe_K4.pdf`[50]

Average profiles of the populations: `Frappe_K4_pops.pdf`[51]

Ranked average profiles of the populations: `Frappe_K4_rankings.pdf`[52]

Here, an 'American' cluster (number 4) is added to the three previous ones: 'African' (cluster 1), 'European' (number 2) and 'East Asian' (number 3).

Compared to the case where $K = 3$, comments regarding the distribution of cluster 1_3 apply also to cluster 1_4. For cluster 2_4, the only notable change with respect to cluster 2_3 is that American populations loose most of their cluster 2 component (this partially affects Mexicans). The same occurs for cluster 3. Altaic, Sino-Tibetan and Hmong-Mien populations also tend to have less cluster 3 proportion, but to a lesser extent, while the opposite tendency is observed for Austronesian, Tai-Kadai and Austro-Asiatic populations. Although it has a somewhat different distribution from cluster 3_3, cluster 3_4 is still the most prominent cluster for South-East, East and North Asia.

Cluster 4 is the main cluster for American populations, particularly for South Americans. Differences between American populations may reflect various degrees of European and African ancestry. In other populations, cluster 4 is rather low, but more present in Altaic populations, Japanese, Koreans and the Sino-Tibetan populations from India (Nysha, Aonaga and Spiti), followed by Hazara, Russians, Pahari, non-Indian Sino-Tibetans, Burusho and Hmong-Mien. It is absent or almost absent in African populations.

Not surprisingly, the profile of Mexicans is approximately composed of half cluster 2 (putative European ancestry) and half cluster 4 (putative American ancestry). The similarity between the Indo-European Hazara and the Altaic Uyghur (see p. 28) is reflected by the fact that Hazara are the Indo-European population with the highest cluster 4 proportion (after Mexicans). The relatively high cluster 4 ranking of Russians might be explained by some degree of admixture with Siberian populations, and that of Pahari by admixture with Sino-Tibetan populations (see p. 28).

To be noted also is the proportion of cluster 4 in Burusho from northern Pakistan, which is similar to that of non-Indian Sino-Tibetan populations, and higher than for the other populations from Pakistan (except Hazara). This population speaks a language isolate which is sometimes grouped with

---

[50] http://dx.doi.org/10.6084/m9.figshare.278
[51] http://dx.doi.org/10.6084/m9.figshare.189
[52] http://dx.doi.org/10.6084/m9.figshare.293

Sino-Tibetan and other languages (including some languages spoken in North America) in a Dene-Caucasian family.

$K = 5$

Raw results: `Frappe_K5.txt`[53]

Profiles of the individuals: `Frappe_K5.pdf`[54]

Average profiles of the populations: `Frappe_K5_pops.pdf`[55]

Ranked average profiles of the populations: `Frappe_K5_rankings.pdf`[56]

Here, there is one cluster for each continent:

- cluster 1, the 'African' cluster (more specifically, 'Sub-Saharan');

- cluster 2, the 'European' cluster;

- cluster 3, the 'Asian' cluster (more specifically, 'East Asian');

- cluster 4, the 'Oceanian' cluster;

- cluster 5, the 'American' cluster;

The distribution of cluster 1_5 is roughly the same as that of cluster 1_4: high in African populations. But some interesting differences can be noticed:
The most conspicuous fact is that cluster 1_5 is almost absent in Oceanian populations, whereas cluster 1_4 represented around 8% of their profile.
A strong decrease is observed in the ANLS populations, who had been previously noticed for the presence of cluster 1_3 (see p. 28). The relative decrease is the strongest for the populations of the Lesser Sunda Islands (Alorese, Kambera, Lamaholot, Lembata, Manggarai), who live the closest to Oceania and for Kensiu (one of the two Malaysian Negrito populations). The decrease is also important for the other Negrito populations (Jehai from Malaysia and Agta, Ati, Ayta and Mamanwa from the Philippines), as well as for the populations of the Andaman Islands.

Apart from those, most populations outside Africa who had at least a few percentage points of cluster 1_4 proportion also have a relatively lower cluster

---

[53] http://dx.doi.org/10.6084/m9.figshare.107
[54] http://dx.doi.org/10.6084/m9.figshare.279
[55] http://dx.doi.org/10.6084/m9.figshare.190
[56] http://dx.doi.org/10.6084/m9.figshare.294

1_5 proportion.

The exceptions to this are Sindhi, Makrani, Balochi, Brahui (from Pakistan), who are affected by a very modest decrease, Siddi and African Americans, who have a negligible decrease, Mexicans, and populations from the Middle East, for which the proportion of cluster 1_5 is even slightly higher than the proportion of cluster 1_4.

This observation might suggest means to distinguish between the genetic signature of recent African ancestry and that pertaining to an ancient out-of-Africa migration. Among populations who had a noticeable cluster 1 for $K = 3$ and $K = 4$, those for which there is no or very little decrease when considering cluster 1_5 probably have recent African ancestry. This is historically known for Siddi and African Americans and probable for Mexicans also. This was hypothesised for Makrani and Sindhi because of the presence of descendants from African slaves or soldiers in the south of Pakistan, and it can be suspected that the same is true for other populations from Pakistan and Middle East. On the contrary, the populations of Oceania and the ANLS mentioned p. do not have known recent African ancestry.

Cluster 2_5 has a distribution very similar to cluster 2_4. But as in the case of cluster 1, cluster 2 almost completely disappears from the profile of Oceanians.

It also almost disappears from the profiles of the Mlabri (Austro-Asiatic hunter-gatherers from northern Thailand) and Manggarai, Lembata, Lamaholot, Kambera and Alorese (Austronesians from the Lesser Sunda Islands). More generally, there is a relative decrease of cluster 2 for Austro-Asiatic and Austronesian populations, as well as for the populations of the Andaman Islands. The decrease also occurs in Jinuo, Karen, and Tai-Kadai populations but is less conspicuous because their cluster 2_4 proportion is already quite low.

At first approximation, cluster 3_4 seems to have been split between cluster 3_5 and cluster 4_5.

Cluster 3_5 is most important in East Asia. Among the populations with a high proportions of cluster 3_5, the rankings according to the importance of this cluster show a tendency for the following gradient:

Chinese and Hmong-Mien > Koreans, Japanese, Taiwanese Austronesians and Tai-Kadai > Tibeto-Burmese, Mon-Khmer, non-Taiwanese Austronesians and Altaic populations.

Among non-Taiwanese Austronesians, the lowest proportions of cluster 3 are observed in the populations of the Lesser Sunda Islands and the Negritos from the Philippines (Ayta, Mamanwa, Agta and Ati).

Among the Mon-Khmer-speaking populations, it is lower for the Malaysian Negritos. It is even lower for the other Austro-Asiatic[57] populations, the Kharia and Santhal from India.
Cluster 3 is also an important component of the profile of the Andamanese populations (Onge and Great Andamanese).

Among Indo-European populations cluster 3 is important in the profiles of Pahari, Hazara and Sahariya. I already mentioned (p. 28) the Altaic ancestry of the Hazara and the proximity between Pahari and Sino-Tibetan populations when discussing their low proportion of cluster 2_3.
Apart from Hazara, Burusho (who speak a language isolate) show a higher cluster 3 proportion than other populations of Pakistan (see also p. 31).
Among Dravidian populations, some Indians from Singapore show an important cluster 3 component. This is probably due to some admixture with Chinese or Malays.

Papuans have almost exclusively cluster 4_5, which also constitutes more than 85% of the profile of Melanesians.
It is an interesting fact that the three first non-Oceanian populations in the ranking according to cluster 4_5 are Alorese, Lembata and Lamaholot, which are also those who are geographically the closest to Papua New Guinea. Apart from populations of the Lesser Sunda Islands, most non-Oceanian populations with a high proportion of cluster 4_5 are either Negritos from Malaysia or the Philippines, Andamanese, or tribal or lower caste populations from India. These populations from India may bear traces of an ancient genetic background, pre-dating the arrival of Dravidian and Indo-European populations.

More generally, cluster 4_5 is an important component for many populations of South and South-East Asia, but it tends to be lower for Sino-Tibetan, Hmong-Mien and Tai-Kadai populations. This distribution is to be related to the gradient observed for cluster 3_5. If we set aside Korean, Japanese and Altaic populations (who have a very low cluster 4_5 proportion) and populations from India and Pakistan (who have a low cluster 3_5 proportion), the distributions of clusters 3_5 and 4_5 are complementary.

Cluster 5_5 has a distribution similar to cluster 4_4, but with a slight increase for most populations of mainland India (the exceptions being Pahari and the Sino-Tibetan Aonaga, Nysha and Spiti), and with a decrease in populations of East and South-East Asia. The populations with the highest proportion

---

[57]Following the classification adopted in Lewis (2009), I divide the Austro-Asiatic populations in two branches: Mon-Khmer (in South-East Asia), and Munda (in India).

of cluster 5_5 are the same as those for cluster 4_4: Americans, followed by Altaic populations.

$K = 6$

Raw results: Frappe_K6.txt[58]

Profiles of the individuals: Frappe_K6.pdf[59]

Average profiles of the populations: Frappe_K6_pops.pdf[60]

Ranked average profiles of the populations: Frappe_K6_rankings.pdf[61]

Here, the 'East Asian' cluster 3_5 is split into a 'northern' component (cluster 3_6) and a 'southern' component (cluster 4_6).

Clusters 1_6 and 2_6 have the same distributions as clusters 1_5 ('African') and 2_5 ('European').

Cluster 3_6 is most important in Japanese and Koreans. The rankings according to this cluster reveal the following (approximate) gradient:
Japanese and Koreans > Altaic and Sino-Tibetans > Hmong-Mien > Tai-Kadai > Mon-Khmer (except Mlabri, Jehai and Kensiu) and Austronesians > Andamanese, Burusho, Munda (Kharia and Santhal) and Dravidians > Indo-Iranian and North American populations.
Other populations have a rather low cluster 3_6 proportion.

Mlabri have almost exclusively cluster 4_6 in their profile. There is a tendency towards the following 4_6 importance gradient:
Mon-Khmer and Austronesians > Tai-Kadai > Hmong-Mien > Sino-Tibetans > Andamanese and Munda > Melanesians > Altaic, Koreans and Japanese.

Among Austronesian populations, cluster 4_6 is lower in the Lesser Sunda Islands and in the Negritos from the Philippines. Among Sino-Tibetan populations, cluster 4_6 is more important in Karen, Lahu and Jinuo, populations sampled near the western Burmese border[62], and less important in Nysha, Aonaga and Spiti, populations sampled in northern India.

Cluster 5_6 has a distribution similar to cluster 4_5 ('Oceanian'), but a significant decrease can be noticed in Austronesian, Mon-Khmer, Tai-Kadai,

---

[58]http://dx.doi.org/10.6084/m9.figshare.108
[59]http://dx.doi.org/10.6084/m9.figshare.280
[60]http://dx.doi.org/10.6084/m9.figshare.191
[61]http://dx.doi.org/10.6084/m9.figshare.295
[62]I will use the abbreviation JKL for this group of populations: Jinuo, Karen, Lahu.

Sino-Tibetan and Hmong-Mien populations. The diversification of the 'East Asian' clusters seems to happen at the expense of the 'Oceanian' cluster.

Cluster 6_6 has a distribution similar to cluster 5_5 ('American'), but with a decrease in Altaic, Japanese, Korean and Sino-Tibetan populations, likely related to the appearance of the 'northern East Asian' cluster 3_6.

$K = 7$

Raw results: `Frappe_K7.txt`[63]

Profiles of the individuals: `Frappe_K7.pdf`[64]

Average profiles of the populations: `Frappe_K7_pops.pdf`[65]

Ranked average profiles of the populations: `Frappe_K7_rankings.pdf`[66]

The new cluster that appears, number 2_7, having its highest frequencies in Dravidian populations, and more generally in India and Pakistan, represents a 'South Asian' tendency. This cluster seems to principally replace parts of the 'European' (2_6) and 'Oceanian' (5_6) clusters.

Cluster 1_7 is mostly unchanged compared to cluster 1_6.

The new cluster 2_7 is almost absent from Africa, Oceania and America. A tiny proportion of the 'European' cluster 2_6 that was detectable in Maya and some African populations has been replaced by cluster 2_7, but cluster 2_6 is mostly preserved as cluster 3_7 in these populations.

The replacement is more visible for populations of Europe and Middle East, except that it does not seem to affect Sardinians, and only very lightly Basques. Populations of Middle East and East Europe are more affected, particularly the Caucasian Adygei.

For the populations of Pakistan, the proportion of the 'Oceanian' cluster (5_6, then 6_7) is greatly reduced. It is replaced by cluster 2_7, which also replaces part of cluster 2_6, so that 2_7 ('South Asian') and 3_7 ('European') are roughly in equal parts. The same observation holds for Altaic populations, but is less conspicuous because clusters 2_6 and 5_6 are less important.

The same is observed also in India, but resulting in a higher $2_7/3_7$ ratio. The proportion of remaining cluster 3_7 is higher in upper-caste Indo-Iranian populations and lower in Andamanese, Munda and Tibeto-Burmese populations.

[63]http://dx.doi.org/10.6084/m9.figshare.109
[64]http://dx.doi.org/10.6084/m9.figshare.281
[65]http://dx.doi.org/10.6084/m9.figshare.192
[66]http://dx.doi.org/10.6084/m9.figshare.296

In East and South-East Asia, 2_6 is mostly replaced by 2_7. The 'Oceanian' component (5_6, then 6_7) is also generally affected by the replacement, but less than in South Asia. Cluster 2_7 highlights the heterogeneity within the Malay and Indian populations from Singapore, probably reflecting the various degrees of Indian ancestry found in the individuals composing these two populations.

The differences in replacement of the 'European' cluster 2_6 by the 'South Asian' cluster 2_6 has the following notable effects on the rankings according to the 'European' cluster (now 3_7):

- an increase of the ranking of Altaic populations (especially Uyghur), Hazara, Fulani and Nilo-Saharan populations (especially Maasai);

- a decrease for Onge, Malaysian Negritos and Munda.

Cluster 4_7 has the same distribution as the 'northern East Asian' cluster 3_6, but with a noticeable increase in proportion and rank for Oceanian populations, Mlabri and Alorese.

Cluster 5_7 has a distribution similar to the 'southern East Asian' cluster 4_6, but with an increase in the rankings for most populations of India and a decrease for Middle East, Europe, Oceania and Japan, and for some Altaic and Nilo-Saharan speakers.

Following the differential replacement of cluster 5_6 by the new 'South Asian' cluster 2_7, the top of the ranking according to the importance of the 'Oceanian' cluster (5_6 then 6_7) becomes clearer:
Papuans have their profile almost exclusively contituted by cluster 6_7, closely followed by Melanesians. Then, populations from the Lesser Sunda Islands have an important cluster 6_7 proportion, which decreases with geographic distance from Papua New Guinea. The decrease continues with Negritos from the Philippines and Andamanese, and then other non-Filipino populations from the Philippines, as well as Toraja from Sulawesi.

Cluster 7_7 has the same distribution as cluster 6_6.

K = 8

Raw results: `Frappe_K8.txt`[67]

---

[67] http://dx.doi.org/10.6084/m9.figshare.110

Profiles of the individuals: `Frappe_K8.pdf`[68]

Average profiles of the populations: `Frappe_K8_pops.pdf`[69]

Ranked average profiles of the populations: `Frappe_K8_rankings.pdf`[70]

Here, a 'non-Niger-Congo' cluster (2_8) replaces parts of the previous 'African' (1_7) and 'European' (3_7) clusters.

Overall, cluster 1_8 has a distribution similar to cluster 1_7. But besides a general decrease in African populations, a contrast can be observed in the variation of rankings in European populations: Sardinians undergo a strong decrease in rankings whereas the rankings of more northern populations (Orcadians, Russians, and to a lesser extant, north Americans of European origins and French) increase.

The new cluster 2_8 constitutes about one third of the profile of the Maasai (who speak a Nilo-Saharan language). It is also present in a significant amount in another Nilo-Saharan-speaking population, the Bulala (but less in the Kaba), and among speakers of Afro-Asiatic languages, particularly in North Africa and Middle East. The Kaba (Nilo-Saharan) and the Hausa (Afro-Asiatic) have little cluster 2_8, like most Niger-Congo-speaking populations

The Niger-Congo-speaking populations with the highest proportion of cluster 2_8 are Bantu from the north-east and Luhya from Kenya (two populations who live in the same region as the Maasai), and the Fulani. This observation may be related to what had been noticed p. 28 when discussing the presence of the 'European' cluster 2_3 in African populations.

Outside Africa and Middle East, cluster 2_8 is above 7% in Italy (including Sardinia), in the Caucasus (Adygei) and in western Pakistan (Makrani, Brahui and Balochi). It would be interesting to include data for more populations of East and North Africa, East Europe and West Asia to get a better view of the geographic distribution of this cluster.

The 'European' cluster 5_8 has roughly the same distribution as cluster 3_7, but is partly replaced by cluster 2_8 in some African populations: Fulani, Maasai, Luhya and Bantu from the north-east, Mada, Kaba and Bulala (where it completely disappears).
This replacement also affects populations from North Africa, Middle East

---

[68]http://dx.doi.org/10.6084/m9.figshare.282
[69]http://dx.doi.org/10.6084/m9.figshare.193
[70]http://dx.doi.org/10.6084/m9.figshare.297

and Italy (including Sardinia), Adygei from the Caucasus, Brahui, Makrani and Balochi from western Pakistan.

The other clusters are mostly unchanged with respect to the case where $K = 7$, with the following correspondences:

Cluster	3_8	4_8	6_8	7_8	8_8
corresponds to cluster	2_7	4_7	5_7	6_7	7_7

$K = 9$

Raw results: `Frappe_K9.txt`[71]

Profiles of the individuals: `Frappe_K9.pdf`[72]

Average profiles of the populations: `Frappe_K9_pops.pdf`[73]

Ranked average profiles of the populations: `Frappe_K9_rankings.pdf`[74]

Here, the 'southern East Asian' cluster which was dominant in Mlabri (6_8) is decomposed in two clusters (6_9 and 7_9). There are now 3 'East Asian' clusters:

- Cluster 4_9 is more present in Altaic, Korean and Japanese populations.

- Cluster 6_9 is more present in Austronesian populations.

- Cluster 7_9 is typical of Malaysian Negritos.

Cluster 4_9 has a similar distribution as cluster 4_8, but with the following changes in the rankings:

- a decrease for Mlabri, Oceanians, and some Austronesian populations;

- an increase for Kensiu (a Malaysian Negrito population), Andamanese, the Himalayan Spiti and Pahari, Srivastata, Hazara, Uyghur, Yakut, Russians, Burusho, North Americans and Colombians.

Cluster 6_9 replaces parts of clusters 4_8 ('northern East Asian') and 6_8 ('southern East Asian'). This replacement most strongly affects Austronesians, but the Negritos from the Philippines and the populations from the Lesser Sunda

---

[71]http://dx.doi.org/10.6084/m9.figshare.111
[72]http://dx.doi.org/10.6084/m9.figshare.283
[73]http://dx.doi.org/10.6084/m9.figshare.194
[74]http://dx.doi.org/10.6084/m9.figshare.298

Islands have less of this cluster than other Austronesians.

Cluster 6_9 is important also in Mon-Khmer (particularly in Mlabri and Ht'in Mal, but not in Malaysian Negritos), Tai-Kadai, Hmong-Mien and Sino-Tibetan populations. Whithin these populations, Tai-Kadai tend to have a higher cluster 6_9 proportion, and Sino-Tibetans tend to have a lower proportion. Cluster 6_9 is found in Koreans, Japanese, Altaic, Melanesians, and some populations of India (most noticeably in Munda).

Cluster 7_9 constitutes a large majority of the profile of Malaysian Negritos. It is found at a significant level in various South and South-East Asian populations, with the populations of the Andaman islands and a majority of Austro-Asiatic speakers among the first populations in the rankings.

Little change occurs for 'African' (1 and 2), 'South Asian' (3), 'Oceanian' (7_8 then 8_9) and 'American' (8_8 then 9_9) clusters, except for a significant decrease in the rankings of Malaysian Negritos.

The 'European' (5) cluster is mostly unchanged, except for a decrease in the rankings of Munda and some Dravidian populations.

K = 10

Raw results: `Frappe_K10.txt`[75]

Profiles of the individuals: `Frappe_K10.pdf`[76]

Average profiles of the populations: `Frappe_K10_pops.pdf`[77]

Ranked average profiles of the populations: `Frappe_K10_rankings.pdf`[78]

Mlabri have now their profile exclusively composed of cluster 7_{10}. This could be due to the low genetic diversity of this population. Indeed, Mlabri seem to have undergone a fairly recent founding effect (Oota et al., 2005).

Cluster 7_{10} partly substitutes the 'Austronesian' and 'southern East Asian' clusters 6_9 (then 6_{10}) and 7_9 (then 8_{10}). This substitution can be evidenced by considering the populations for which the decreases in the 'Austronesian' and 'southern East Asian' clusters are the highest.

Decrease in the 'Austronesian' cluster:

- more than 8 points for Mlabri, Ht'in Mal;

---

[75]http://dx.doi.org/10.6084/m9.figshare.112
[76]http://dx.doi.org/10.6084/m9.figshare.284
[77]http://dx.doi.org/10.6084/m9.figshare.195
[78]http://dx.doi.org/10.6084/m9.figshare.299

- more than 7 points for Temuans;

- more than 6 points for Plang Blang, Wa;

- more than 5 points for Jinuo, Karen, Cambodians, Lawa, Palaung;

- more than 4 points for Bidayuh, Dayak, Javanese, Sunda, Tai Yuan;

- more than 3 points for Aonaga, Nysha, Lahu, Santhal, Mon, Malays from Singapore, Dai, Tai Khuen, Tai Yong, Tai Lue, Zhuang;

- more than 2 points for Satnami, Kharia, Hmong, Iu Mien, Ayta, Malays, Hakka, Tujia, Jiamao.

Decrease in the 'southern East Asian' cluster:

- more than 5 points for Malbri;

- more than 4 points for Ht'in Mal;

- more than 3 points for Temuans, Plang Blang, Wa;

- more than 2 points for Pedi, Javanese, Sunda, Jinuo, Karen, Cambodians, Lawa, Palaung.

This is correlated with the head of the rankings according to the importance of cluster 7_{10}.

Apart from the Mlabri, whose case has been already discussed, the populations with the highest proportions of cluster 7_{10} are the other non-Negrito Mon-Khmer populations (Ht'in Mal, Plang Blang, Wa, Lawa, Cambodians, Palaung, Mon), the Tibeto-Burmese populations sampled near the Burmese border (JKL, see p. 34), the Tai-Kadai populations, and the Austronesian populations from the Malaysian peninsula, Java and Borneo.

Except for the decreases mentioned above, the distribution of clusters 6_{10} and 8_{10} are fairly similar to those of clusters 6_9 ('Austronesian') and 7_9 ('Malaysian Negrito') respectively.

The other clusters are mostly unchanged with respect to the case where $K = 9$, with the following correspondences:

Cluster	1_{10}	2_{10}	3_{10}	4_{10}	5_{10}	9_{10}	10_{10}
corresponds to cluster	1_9	2_9	3_9	4_9	5_9	8_9	9_9

K = 11

Raw results: `Frappe_K11.txt`[79]

Profiles of the individuals: `Frappe_K11.pdf`[80]

Average profiles of the populations: `Frappe_K11_pops.pdf`[81]

Ranked average profiles of the populations: `Frappe_K11_rankings.pdf`[82]

The 'African' putative ancestry is now divided in 3 clusters. A new 'Khoisan-Pygmy' cluster is added to the previously identified 'general Sub-Saharan' and 'East African-West Asian' clusters.

Cluster 1 ('general Sub-Saharan') undergoes an important decrease in Pygmies and San (more than 40 percentage points). A decrease is also observable in other African populations, most notably in south-eastern Bantu populations (Pedi, Tswana, Xhosa, Sotho, Zulu).
Outside Africa, a decrease in cluster 1 is noticeable in Negritos from the Philippines.

Cluster 2_{11} is present mainly in African populations. It reaches its highest proportions in Mbuti Pygmies (72.10%), San (67.58%) and Biaka Pygmies (52.24%). The next populations according to the importance of this cluster are Bantu populations from south-eastern Africa (Pedi, Tswana, Xhosa, Sotho, Zulu). This is probably a consequence of genetic exchanges between Khoisan and Bantu populations in this region (see Schuster et al., 2010).
It should be noticed that, in the rankings according to cluster 2_{11}, the first two populations without obvious African origins are Ayta and Agta, two of the populations mentioned p. 28 about a possible genetic trace of an early out-of-Africa migration in the populations of maritime South-East Asia.
It may be interesting in this regard to consider the proportion of cluster 2_{11} with respect to the total of the three 'African' clusters 1_{11}, 2_{11} and 3_{11}:

Populations from the Lesser Sunda Islands:

- Kambera 76.26%

- Lamaholot 59.89%

- Manggarai 55.46%

---

[79] http://dx.doi.org/10.6084/m9.figshare.113
[80] http://dx.doi.org/10.6084/m9.figshare.285
[81] http://dx.doi.org/10.6084/m9.figshare.196
[82] http://dx.doi.org/10.6084/m9.figshare.300

- Lembata 50.13%

- Alorese 31.35%

Negritos from the Philippines:

- Ayta 78.39%

- Agta 65.64%

- Mamanwa 57.32%

- Ati 49.54%

Malaysian Negritos:

- Jehai 77.11%

- Kensiu 23.60%

Andamanese:

- Onge 42.31%

- Great Andamanese 11.84%

Known Sub-Saharan ancestry in historical times (through African slaves or soldiers):

- Siddi 9.14%

- African Americans 6.41%

Probable Sub-Saharan ancestry (same reasons as above, at least for some individuals):

- Sindhi 15.30%

- Makrani 12.63%

Possible Sub-Saharan ancestry (through African slaves or soldiers, or because of geographical proximity with the above-mentioned populations):

- Mexicans 20.46%

- Brahui 10.27%

- Balochi 9.69%

- Palestinians 6.16%

- Druze 6.09%

- Bedouins 3.23%

- Mozabites 4.33%

Bantu populations from southern Africa (possible Khoisan ancestry):

- Pedi 26.03%

- Tswana 25.37%

- Xhosa 19.90%

- Sotho 19.19%

- Zulu 15.11%

- Herero 9.88%

- Ovambo 4.02%

Khoisan and Pygmies:

- Mbuti Pygmies 72.27%

- San 68.24%

- Biaka Pygmies 52.66%

The other Sub-Saharan populations have this proportion ranging from 2.59% (Yoruba) to 11.71% (Maasai). This proportion cannot be reasonably evaluated in Papuans and Melanesians because the cumulated proportion of their profile representing putative African ancestry is too low (one Melanesian sample is at 99.99% and the other at 0.58%, but they are both supposed to be taken from the same population).

Except for Great Andamanese and Kensiu, the populations previously hypothesized to bear the trace of an ancient out-of-Africa migration (ANLS) have more than 30% of their total 'African ancestry' represented by cluster 2_{11}. Among African populations or populations with known or suspected African ancestry, only Pygmies and San have this proportion higher than 30%. Great Andamanese and Kensiu still have a higher relative proportion of cluster 2_{11} than the Sub-Saharan populations without suspected Khoisan admixture.

This suggests a scenario in which one or more populations from the same stock as Khoisan and Pygmies migrated to South-East Asia, and that the Negritos from Malaysia and the Philippines and the populations of the Andaman and Lesser Sunda Islands are partially descendants of these populations.

The observations on the variations in the 'African' cluster when the 'Oceanian' cluster first appeared may be related to this (see p.).

Cluster 3_{11} corresponds to cluster 2_{10}, but there is a tendency for the rankings of San, Pygmies, south-eastern Bantu and ANLS populations to decrease.

The other clusters are mostly unchanged with respect to the case where $K = 10$, with the following correspondences:

Cluster	4_{11}	5_{11}	6_{11}	7_{11}	8_{11}	9_{11}	10_{11}	11_{11}
corresponds to cluster	3_{10}	4_{10}	5_{10}	6_{10}	7_{10}	8_{10}	9_{10}	10_{10}

K = 12

Raw results: `Frappe_K12.txt`[83]

Profiles of the individuals: `Frappe_K12.pdf`[84]

Average profiles of the populations: `Frappe_K12_pops.pdf`[85]

Ranked average profiles of the populations: `Frappe_K12_rankings.pdf`[86]

The 'Khoisan-Pygmy' cluster disappears. The comparisons shall therefore be made with the situation at $K = 10$.

A rearrangement of the 'East Asian' clusters occurs:

---

[83] http://dx.doi.org/10.6084/m9.figshare.114
[84] http://dx.doi.org/10.6084/m9.figshare.286
[85] http://dx.doi.org/10.6084/m9.figshare.197
[86] http://dx.doi.org/10.6084/m9.figshare.301

- There are 2 'Austronesian' clusters (6_{12} and 7_{12}), one of which (6_{12}) is in fact more specific to the non-Filipino populations of the Philippines. Cluster 7_{12} has a reinforced Austronesian character.

- A 'continental South-East Asian' cluster appears.

- The 'northern East Asian' cluster 4 acquires a more 'maritime' aspect.

- The 'Mlabri-specific' and 'Malaysian Negrito-specific' clusters are maintained.

The 'African' clusters 1 and 2 and the 'European' cluster 5 do not change much, the most notable difference with respect to the case where $K = 10$ is a decrease in the rankings for Mamanwa.

The distribution of the 'Indian' cluster 3 is mostly unchanged. A tendency towards a decrease in the rankings can be observed for the populations of the Philippines (especially in Mamanwa), Taiwan and Japan.

The 'northern East Asian' cluster 4 undergoes a significant decreases in many Asian populations: Sino-Tibetans, Hmong-Mien, Mon-Khmer (except Mlabri and Malaysian Negritos), Altaic populations, Pahari, Koreans, Tai-Kadai, Hazara, Japanese, Sahariya. Among these populations, the decrease tends to be lower in Japanese, Tai-Kadai and southern Chinese populations. Cluster 4 increases in some Austronesian populations. These differences lead to an increased contrast between populations of Japan and the other populations of northern East Asia. The rankings of Filipinos and Austronesian Taiwanese increase.

Cluster 6_{12} represents about two thirds of the profile of Mamanwa, nomadic Negritos from the Philippines living in the north of Mindanao. It also represents more than 8% of the profiles of the other non-Filipino populations of the Philippines (Ati, Ayta, Agta, Iraya, Manobo).

Cluster 7_{12} corresponds to the 'Austronesian' cluster 6_{10}, but with significant changes. A decrease is observed for many populations of Central and East Asia. The decrease in percentage points is more important in Mamanwa, Hmong-Mien, Mon-Khmer (except Mlabri and Malaysian Negritos), JKL and Tai-Kadai. This decrease is still significant in populations in which the proportion of cluster 6_{10} was not very high. This results in a strong relative decrease for the Sino-Tibetan populations of India (Aonaga, Nysha and Spiti), Pahari, Kashmiri, Hazara, and Altaic populations. An increase can be noted in Okinawans. These variations reveal a contrast between 'continental' and 'maritime' populations.

The Austronesian populations are more grouped in the top of the rankings according to cluster 7_{12} than they were for cluster 6_{10}: The first 21 positions are occupied by Austronesian populations, and they are all found in the 38 first positions. Tai-Kadai are the second group of populations according to the importance of cluster 7_{12}. They rank between 22 and 34. It should be noted in this regard that it has been proposed that Tai-Kadai languages are part of the Austronesian family (Sagart, 2004). Cambodians are the non-Austronesian and non-Tai-Kadai population with the highest proportion of cluster 7_{12}. This could be explained by a possible admixture with Cham, an Austronesian population which once occupied part of southern Indochina, and which is still present in Cambodia, or even by the presence of Cham people in the Cambodian sample.

Cluster 8_{12} is similar to the 'Mlabri-specific' cluster 7_{10}, but with an notable relative decrease for Hmong-Mien, Pahari and Tibeto-Burmese from continental south China (Naxi, Yizu, Lahu) and north-east India (Aonaga, Nysha).

Cluster 9_{12} corresponds to the 'Malaysian Negrito-specific' cluster 8_{10}, but with an important rank decrease for Mamanwa.

Cluster 10_{12} constitutes an important proportion of the profiles of populations of East Asia. The following approximate cluster 10_{12} gradient shows a 'southern continental' > 'northern maritime' tendency within East Asia:
Hmong-Mien (except She), Tibeto-Burmese (except Spiti) and Palaungic (Lawa, Palaung, Wa, Plang Blang) Mon-Khmer > Ht'in Mal and Tai-Kadai (except Zhuang) > She, Chinese and Zhuang > Mon, Cambodians, Tungusic (Hezhen, Xibo, Oroqen) and Mongolic (Tu, Mongola, Daur) Altaic, Pahari, Spiti and Koreans > Austronesian populations of Java, the Malaysian peninsula and Borneo, Turkic (Yakut and Uyghur) Altaic, Hazara, Sahariya and Japanese.

Cluster 11_{12} corresponds to the 'Oceanian' cluster 9_{10}, but with a decrease for Negritos from the Philippines and important rank decreases in some populations of Sumatra, Taiwan, the Philippines, and Japan.

Cluster 12_{12} corresponds to the 'American' cluster 10_{10}. A decrease occurs for Ami and Atayal from Taiwan and Mamanwa and Iraya from the Philippines.

K = 13

Raw results: `Frappe_K13.txt`[87]

Profiles of the individuals: `Frappe_K13.pdf`[88]

Average profiles of the populations: `Frappe_K13_pops.pdf`[89]

Ranked average profiles of the populations: `Frappe_K13_rankings.pdf`[90]

At $K = 13$, there are several important changes:

- The 'Khoisan-Pygmy' cluster observed at $K = 11$ reappears (2_{11} then 2_{13}).

- A new 'Middle Eastern' cluster (4_{13}) appears.

- The cluster specific to the Negritos from the Philippines (6_{12}) disappears.

The results shall thus be compared to the situation at $K = 11$.

Cluster 1_{13} corresponds to cluster 1_{11}. It decreases in African populations, particularly in the Nilo-Saharan-speaking Maasai and Bulala, but also in Kaba (who also are Nilo-Saharan speakers), and in the two East African Niger-Congo populations Luhya and Bantu from the north-east (see p. 28), as well as in the Afro-Asiatic Mada. A less important decrease occurs for the Onge from the Andaman Islands, but this leads to a very strong effect in terms of relative decrease and rankings.

Cluster 2_{13} corresponds to cluster 2_{11}. An important rank decrease can be noted in Vaish, Onge, Russians and Kamsali, and an increase in Druze.

Cluster 3_{13} roughly corresponds to cluster 3_{11} (it is present mainly in East and North Africa and Middle East) but is now less important in populations from West Asia, North Africa and Europe.
The 'Sub-Saharan' character of cluster 3_{13} is reinforced with respect to cluster 3_{11} because important decreases occur for many populations, particularly in Middle East, North Africa, Europe (especially in Sardinia, southern Italy and in the Caucasus), and Pakistan. Simultaneously, most Sub-Saharan populations undergo an increase in cluster 2. Notable exceptions are Zulu and Ovambo, two Bantu populations from southern Africa, and Fulani, for which there is a notable decrease.

The new 'Middle Eastern' cluster (4_{13}) constitutes about one third of the profiles of the populations of Middle East. It is also important for the populations of western Pakistan (Brahui, Makrani and Balochi), the Adygei

---

[88] http://dx.doi.org/10.6084/m9.figshare.287
[89] http://dx.doi.org/10.6084/m9.figshare.198
[90] http://dx.doi.org/10.6084/m9.figshare.302

(Caucasus), the Mozabites (North Africa), and the Kalash (more than 15% in these populations). It is also present at a significant level in the other populations of Pakistan, in Kashmiri and in the populations of Italy (including Sardinia),

Cluster 5_{13} corresponds to the 'South Asian' cluster 3_{11}. A slight increase can be noted in West and North European populations.

Cluster 6_{13} corresponds to the 'northern East Asian' cluster 4_{11}. A decrease occurs in southern and continental populations. The decrease has the following approximate importance gradient:
Tibeto-Burmese and Palaungic Mon-Khmer > Altaic (except Uyghur), Pahari, Ht'in Mal, Hmong-Mien > Mon, Chinese and Koreans > Tai-Kadai and Cambodians > populations of Japan, Hazara and Uyghur > populations of Java.

Cluster 7_{13} corresponds to the 'European' cluster 6_{11}. A general decrease is observed, which is more important in populations from the Middle East (more than 12 percentage points lost in these populations). The contrast between non-Caucasian Europeans and other populations is reinforced because the new cluster 4_{13} replaces a more important part of the 'European' cluster in Adygei and populations from Middle East, North Africa and Pakistan than in non-Caucasian European populations. Non-Caucasian Europeans have more than 67% cluster 7_{13}, the Adygei are at 51.7%, and the other populations are below 50%. The proportion of the 'European' cluster remains above 20% in Middle East, North Africa and Pakistan, as well as in Kashmiri, Uyghur and Mexicans.

Cluster 8_{13} is similar to the 'Austronesian' cluster 7_{11}, but with a significant decrease in many populations of East Asia, most notably in Mon-Khmer (except Malaysian Negritos and Mlabri), Sino-Tibetans (except Spiti), Austronesian populations of Java, Borneo and the Malaysian peninsula, Tai-Kadai and Hmong-Mien. Within these populations the following contrasts can be noted:

- Among Mon-Khmer populations, the decrease is stronger in Ht'in Mal and Palaungic.

- Among Sino-Tibetans, the decrease is stronger in non-Spiti Tibeto-Burmese, especially in JKL, and less important in northern Chinese.

- Among Tai-Kadai, the decrease is slightly less strong in the eastern populations (Jiamao and Zhuang).

- Among Hmong-Mien, the decrease is less strong in She.

A slight increase occurs in Onge and Mamanwa.

The decreases in the 'Austronesian' cluster correlate quite well with the appearance of a 'general southern East Asian' cluster (9_{13}). This cluster accounts for almost one third of the profiles of Palaungic, Ht'in Mal, and JKL populations. It is present at more than 7% in Austro-Asiatic (except Malbri and the Kensiu Malaysian Negritos), Hmong-Mien, Sino-Tibetans, Tai-Kadai, Austronesians from Java, Borneo, the Malaysian peninsula and Sumatra (except Mentawai), Altaic, Koreans, Pahari and Sahariya. Contrasts similar as above are visible:

- Cluster 9_{13} is more important in Palaungic and Ht'in Mal than in the other Mon-Khmer populations.

- Among Sino-Tibetans, it is more important in non-Spiti Tibeto-Burmese (especially in JKL) than in Chinese, and it is less important in Spiti.

- Among Tai-Kadai, it is more important in western populations.

- Among Hmong-Mien, it is less important in She.

- The importance of cluster 9_{13} is quite variable within Austronesian populations. It is more important in Temuans (from the Malaysian peninsula) and in the populations of Java.

- Among Altaic populations, it is less important in the Turkic Yakut and Uyghur.

Cluster 10_{13} corresponds to the 'Mlabri-specific' cluster 8_{11}. A decrease can be observed, which also correlates with the appearance of cluster 9_{13}. It is stronger in Ht'in Mal and Palaungic Mon-Khmer, JKL and Temuans (more than 2.5 percentage points).
A slight increase can be noticed in some populations of Taiwan and the Philippines, in Japan and in Mentawai.

Cluster 11_{13} corresponds to the 'Malaysian Negrito' cluster 9_{11}. In a similar way as above, a decrease occurs in the populations that have an important proportion of cluster 9_{13}, particularly in Ht'in Mal and Palaungic Mon-Khmer, JKL, Temuans, Bidayuh (from Borneo) and the populations of Java (more than 4.5 percentage points).

An increase occurs in populations of Japan, the Philippines, Taiwan, Sulawesi and in Mentawai.

Cluster 12_{13} corresponds to the 'Oceanian' cluster 10_{11}. A decrease occurs in many Austronesian populations (particularly in the Philippines, less in Java), in Melanesians, Onge and Okinawans. The rankings of Taiwanese Austronesians and Mentawai strongly decreases.

Cluster 13_{13} corresponds to the 'American' cluster 11_{11}. A decrease occurs in Ami from Taiwan and in Indo-European (except Pahari and populations from Pakistan), Dravidian (except Brahui from Pakistan) and Munda populations.

K = 14

Raw results: `Frappe_K14.txt`[91]

Profiles of the individuals: `Frappe_K14.pdf`[92]

Average profiles of the populations: `Frappe_K14_pops.pdf`[93]

Ranked average profiles of the populations: `Frappe_K14_rankings.pdf`[94]

The 'Middle Eastern' cluster disappears, but the 'Khoisan-Pygmy' cluster is still there. Therefore, for the 'African' clusters, the comparisons will be made with the situation at $K = 11$, which is probably quite similar.

The Asian clusters are highly reorganized:

- There are two 'Austronesian' clusters. Cluster 7_{14} is dominant in Borneo, Java and the Malaysian peninsula and cluster 8_{14} is dominant in the Philippines.

- There is a 'southern East Asian' cluster (11_{14}) predominant in Hmong-Mien and Sino-Tibetan populations.

- There is a cluster specific to the Andamanese and Negritos from the Philippines (12_{14}).

- The 'Indian' (4_{14}), 'northern East Asian' (5_{14}), 'Mlabri-specific' (9_{14}), and 'Malaysian Negrito' (10_{14}) clusters can still be identified.

[91] http://dx.doi.org/10.6084/m9.figshare.116
[92] http://dx.doi.org/10.6084/m9.figshare.287
[93] http://dx.doi.org/10.6084/m9.figshare.199
[94] http://dx.doi.org/10.6084/m9.figshare.303

Cluster 1_{14} corresponds to the 'general Sub-Saharan' cluster 1_{11}. The only important difference is that it disappears from the profile of Onge.

Cluster 2_{14} is similar to the 'Khoisan-Pygmy' cluster 2_{11}. It disappears from the profile of Onge and decreases in Great Andamanese, in the Negritos from the Philippines and in some populations of India.

Cluster 3_{14} corresponds to the 'East African-West Asian' cluster 3_{11}. It disappears from the profile of Onge, and also slightly decreases in Great Andamanese, Sardinians, and in the populations of Middle East and North Africa.

Cluster 4_{14} is similar to the previously described 'Indian' cluster. It constitutes the majority of the profiles of most Dravidian populations. The exceptions are Brahui from Pakistan (38.99%) and the 'African Indians' Siddi (16.21%). It can be noted that the Indians from Singapore have a somewhat lower cluster 4_{14} proportion compared to the Dravidian populations of India. This could be explained by some admixture with Chinese or Malay populations.
Cluster 4_{14} is also important in other populations of India and Pakistan. It is above 50% in the Indo-Iranian populations of India except Sahariya (48.47%), Kashmiri (45.41%) and Pahari (27.09%). It is important in Munda and still notable in Great Andamanese and Spiti. In Pakistan the proportion of cluster 4_{14} is highest in Sindhi (44.68%) and lowest in Hazara (17.39%). Outside Pakistan and India, cluster 4_{14} is notable in Adygei, Uyghur and Mon. This presence in Mon could be related to the long time period when Indochina received commercial, political and cultural inputs from India and Sri Lanka (see also p. 27).

Cluster 5_{14} is similar to the previously described 'northern East Asian' cluster. However, it displays a clear contrast between the populations of Japan and the other populations. This seems stronger than the contrast already observed at $K = 12$. Cluster 5_{14} constitutes almost 75% of the profile of Okinawans, almost 65% in Japanese and almost 50% in Koreans. It then decreases according to the following approximate gradient:
Altaic (except Uyghur) > Sino-Tibetans (except Spiti, southern Chinese and JKL)> southern Chinese, Spiti, She, Hazara, Uyghur and Pahari> JKL, Miaozu, Iu Mien, Palaungic, Mon, Cambodians, Filipinos and Austronesian Taiwanese.

Cluster 6_{14} is similar to the previously identified 'European' cluster, except for an important decrease in the rankings of San and Pygmies and an increase in the rankings of Mamanwa, it's distribution resembles much that observed at $K = 12$ (cluster 5_{12}).

Cluster 7_{14} is a 'South-East Asian' cluster, most predominant in Bidayuh from Borneo. It is present at a notable level in Austronesian populations (except those from Taiwan and the Philippines), some Austro-Asiatic populations, JKL and Tai-Kadai.

Among Austronesians, it is more important in the populations of Borneo (Bidayuh and Dayak), Java (Javanese and Sunda) and the Malaysian peninsula (Temuans, Malays) and much less important in some non-Filipino populations of the Philippines. Among Austro-Asiatic, it is more important in Ht'in Mal and Palaungic and very low in Kensiu Negritos and Mlabri. Among Tai-Kadai, it is less important in the eastern populations (Zhuang and Jiamao).

Cluster 8_{14} is another 'Austronesian' cluster, which is somewhat complementary to the previous one. It is most important in the Philippines, Taiwan, Sulawesi (Toraja) and Sumatra (Mentawai, Batak and Malays[95]). It is present at a notable level in Tai-Kadai, Chinese and Hmong-Mien. Among Tai-Kadai, it is more important in the eastern populations, and among Chinese, it is less important in northern populations. Cluster 8_{14} is also present in other Sino-Tibetan populations, but at lower levels, and in Cambodians, Mon, Japanese, Koreans and Melanesians.

Cluster 9_{14} corresponds to the 'Mlabri-specific' cluster previously identified. It constitutes almost entirely the profile of Mlabri. It is slightly above 9% in Ht'in Mal, slightly above 7% in Temuans and is otherwise present at a low level in various populations of South-East Asia.

Cluster 10_{14} corresponds to the 'Malaysian Negrito' cluster previously identified, but with the notable difference that it disappears from the profile of Onge. It also decreases in Great Andamanese, the Austronesian populations of Java, Borneo and the Malaysian peninsula, Austro-Asiatic (except Mlabri) and JKL populations.

Cluster 11_{14} is a 'southern East Asian' cluster somewhat similar to cluster 10_{12}. Like cluster 10_{12}, it has its highest proportion in Hmong, but there are significant differences. Decreases are observed in Austro-Asiatic populations (except Malbri and Kensiu), Austronesian populations of Java, Borneo, and peninsular Malaysia, and JKL. It increases with respect to cluster 10_{12} in Taiwanese Austonesians, Hmong, She, Chinese, Tujia and the eastern Tai-Kadai Jiamao (more than 6.5 percentage points), and to a lesser extent in Koreans, Japanese, Altaic, the other Hmong-Mien, Tai-Kadai and Tibeto-Burmese populations (except JKL), the populations of the Philippines, Sulawesi and Mentawai, Hazara and Pahari.

---

[95]The Malay individuals were sampled in both peninsular Malaysia and Sumatra.

Cluster 12_{14} is specific to Andamanese populations and Negritos from the Philippines. It constitutes almost entirely the profile of Onge, and more than one third of that of Great Andamanese. It is quite important in the profiles of Negritos from the Philippines and is notable in some populations of India (particularly Dravidian, tribal or lower caste populations).

Cluster 13_{14} is similar to the previously identified 'Oceanian' cluster, but almost disappears from Onge and is halved in Great Andamanese. A significant rank decrease can be noticed in Okinawans, Srivastava and Vaish.

Cluster 14_{14} is similar to the 'American' cluster previously identified, with a strong relative decrease in Onge.

K = 15

Raw results: `Frappe_K15.txt`[96]

Profiles of the individuals: `Frappe_K15.pdf`[97]

Average profiles of the populations: `Frappe_K15_pops.pdf`[98]

Ranked average profiles of the populations: `Frappe_K15_rankings.pdf`[99]

At $K = 15$, a 'Middle Eastern' cluster is present, as was the case at $K = 13$. The other clusters correspond to those present at $K = 14$.

Clusters 1_{15} and 2_{15} are much similar to the 'general Sub-Saharan' cluster 1_{13} and the 'Khoisan-Pygmy' cluster 2_{13} respectively, except for an important rank decrease for Onge.

Cluster 3_{15} is much similar to the 'East African' cluster 3_{13} except for an important rank decrease for Onge and Great Andamanese.

Cluster 4_{15} is similar to cluster 4_{13}, in as much as it constitutes about one third of the profiles of the populations of Middle East. But there are otherwise important differences. It decreases in many populations of Pakistan and India, as well as in some populations of the Philippines and in Uyghur. It is reinforced in Middle East, Italy, North Africa, Maasai and Fulani.

Cluster 5_{15} is similar to the 'Indian' cluster previously identified. Compared to 5_{13}, an important decrease occurs in Great Andamanese, it disappears

---

[96]http://dx.doi.org/10.6084/m9.figshare.117
[97]http://dx.doi.org/10.6084/m9.figshare.289
[98]http://dx.doi.org/10.6084/m9.figshare.200
[99]http://dx.doi.org/10.6084/m9.figshare.304

from Onge, and increases in Middle East and western populations of Pakistan. Compared to 4_{14}, a decrease occurs for Brahui and Middle Eastern populations and a slight increase for populations of West and North Europe.

Cluster 7_{15} is similar to the 'European' cluster 7_{13}. There is a decrease in populations from Middle East, Italy, Caucasus and western Pakistan, and an increase in Kalash.

The other clusters are mostly unchanged with respect to the case where $K = 14$, with the following correspondences:

Cluster	6_{15}	8_{15}	9_{15}	10_{15}	11_{15}	12_{15}	13_{15}	14_{15}	15_{15}
corresponds to cluster	5_{14}	7_{14}	8_{14}	9_{14}	10_{14}	11_{14}	12_{14}	13_{14}	14_{14}

K = 16

Raw results: `Frappe_K16.txt`[100]

Profiles of the individuals: `Frappe_K16.pdf`[101]

Average profiles of the populations: `Frappe_K16_pops.pdf`[102]

Ranked average profiles of the populations: `Frappe_K16_rankings.pdf`[103]

At $K = 16$, the cluster specific to the Andamanese populations again disappears. The 'Austronesian' clusters are reorganized, with the appearance of a cluster specific to the non-Filipino populations of the Philippines (10_{16}), as was the case at $K = 12$. The 'American' cluster is now separated in a 'northern' cluster (15_{16}) and a 'southern' cluster (16_{16}).
The most important changes observed in the other clusters are related to the above-mentioned cluster appearances and disappearances: They often affect Andamanese, Negritos from the Philippines and populations from North America.

Clusters 1_{16} and 2_{16} correspond to clusters 1_{15} and 2_{15} respectively, except for an important rank decrease in Mamanwa and an important rank increase in Onge.

Cluster 3_{16} corresponds to cluster 3_{15}, except for important rank decreases in Kalash, Pima and Mamanwa, and important rank increases in Onge, Great Andamanese, Ayta and Ovambo.

---

[100] http://dx.doi.org/10.6084/m9.figshare.118
[101] http://dx.doi.org/10.6084/m9.figshare.290
[102] http://dx.doi.org/10.6084/m9.figshare.201
[103] http://dx.doi.org/10.6084/m9.figshare.305

Cluster 4_{16} is similar to cluster 4_{15}. A significant increase occurs in many populations where cluster 4_{15} was already important (West Asia, North Africa, Europe). An important rank decrease occurs for Mamanwa, Ati, Ayta, Pima, Ovambo, Pedi and Great Andamanese.

Cluster 5_{16} is similar to cluster 5_{15}. An increase occurs in Andamanese, in some tribal and lower caste populations of continental India and in some Negritos from the Philippines (Ayta, Agta and Ati). This increase is particularly important for Onge. An important rank decrease is observed for Mamanwa and Pima.

Cluster 6_{16} is similar to cluster 7_{15}. A decrease occurs in the populations of Middle East, North Africa, Caucasus, Italy (more in Sardinia, less in the north) and western Pakistan. The decreases somewhat reflect the increases observed for the 'Middle Eastern' cluster. Important rank decreases affect Pima, the Negritos from the Philippines Mamanwa and Agta, and some populations of Southern Africa (Herero, Tswana and San), and important rank increases are observed for Onge and Ayta.

Cluster 7_{16} is similar to the 'northern East Asian' cluster 6_{15}. Increases occur for Andamanese and for the Negritos from the Philippines Ayta, Ati and Agta. This increase is particularly important for Onge. The North American populations Pima and Maya and the North Asian population Yakut lose more than 2 percentage points. A decrease is also observed for Colombians, Oroqen and Hezhen. For American populations, the decrease in the 'northern East Asian' cluster manifests itself also by an important rank decrease. It is interesting to note that Yakut and Oroqen are the two northernmost populations of the dataset. This variation correlation between northern East Asian and North American populations might reflect some common ancestry, either dating back from the colonization of America, either due to later exchanges.

Cluster 8_{16} roughly corresponds to the 'Taiwan-Philippine Austronesian' cluster 9_{15}. Compared to cluster 9_{15}, an important decrease affects the Negritos from the Philippines. A significant increase occurs in Hmong-Mien, Tai-Kadai, southern Chinese and Taiwanese Austronesians. Cluster 8_{16} is thus most important in Taiwanese populations, followed by Mentawai and the non-Negrito populations of the Philippines.

Cluster 9_{16} is similar to cluster 8_{15}. An important rank decrease affects Taiwanese Austronesians and Pima, and an important rank increase is observed in Onge, Ayta and Mamanwa. In Onge, this corresponds to a significant increase in percentage points.

Similarly to cluster 6_{12}, cluster 10_{16} is dominant in Mamanwa and important

in the other non-Filipino populations of the Philippines. However, it has a higher level in these populations, as well as in Andamanese and in many Austronesian and Austro-Asiatic populations. It is much lower in northern and western European populations as well as in Kalash.

Cluster 11_{16} corresponds to the 'Mlabri-specific' cluster 10_{15}, with a slight increase in Andamanese and in the Austronesian populations of Taiwan, and a slight decrease in Mamanwa and Pedi.

Cluster 12_{16} corresponds to the 'Malaysian Negrito-specific' cluster 11_{15}, with an important increase in Onge, and an important rank decrease in Mamanwa.

Cluster 13_{16} is similar to cluster 12_{15}, but with a decrease in some southern East Asian populations, particularly in Hmong-Mien, Southern Chinese, Tai-Kadai, and Taiwanese Austronesians. Among Tai-Kadai, the decrease is stronger in eastern populations. The distribution of cluster 13_{16} is thus slightly 'flattened' with respect to that of cluster 12_{15}. Important rank decreases can be noticed for Pima, Onge and Agta.

Cluster 14_{16} corresponds to cluster 14_{15}, except for a strong decrease in Mamanwa and a strong increase in Onge.

Cluster 15_{16} is a 'northern American' cluster. It constitutes almost 75% of the profile of Pima (Mexico), which is the northernmost native American population in the dataset, and almost 30% for Maya and Colombians. It is a notable component of the profile of the Mexicans sampled in Los Angeles. Apart from these populations, it is only present at a low level, principally in some Indo-European and Altaic populations, in Burusho and in Spiti.

Cluster 16_{16} is a 'southern American' cluster. It constitutes almost entirely the profiles of the Tupi-speaking Amazonian populations (Surui and Karitiana). It is important in the other American populations and decreases according to a south > north gradient. Outside America, it is below 5% except in Yakut (7.39%) and Oroqen (5.34%), which are the two northernmost populations of the dataset. This may be related to the decrease observed for cluster 7_{16} with respect to cluster 6_{15}.